A SAS® Macro for Calculating Bootstrapped Confidence Intervals About a Kappa Coefficient

Robert A. Vierkant, Marshfield Medical Research Foundation, Marshfield, WI

ABSTRACT

Cohen's kappa coefficient has become a standard method for measuring the degree of agreement between two raters. Confidence intervals for kappa and weighted kappa based on its asymptotic variance are available in the SAS system through the FREQ procedure. However, this variance can become unreliable as sample size decreases or as kappa approaches unity. This paper presents a SAS macro for calculating confidence intervals about kappa or weighted kappa using bootstrap resampling methodology, and is intended for an audience with a basic understanding of statistics.

INTRODUCTION

The kappa coefficient is a standard measure of interrater agreement for categorical data (Cohen, 1960). Kappa was originally designed for use in psychological studies, but its implementation has expanded to many other fields, including biology, ergonomics, and epidemiology, among others. Cohen (1968) also developed a weighted kappa coefficient to be used when the degree of disagreement between two raters can be quantified. The weighted kappa is a generalization of the simple kappa coefficient, using weights to effectively assign partial credit to near, but not exact, agreement. Fleiss and others (1969) estimated the variance of kappa based on the asymptotic normality of its sampling distribution. This variance, used by SAS procedure PROC FREQ (1996) to calculate confidence intervals for kappa, works well when the number of subjects is large relative to the number of categories, but has been proven unreliable otherwise, especially as kappa approaches unity (Fleiss and Cicchetti, 1978). This paper presents an alternative method of calculating confidence intervals about kappa or weighted kappa using SAS macros and bootstrap resampling methodology (Efron, 1982). An example is presented that compares the bootstrapped confidence intervals with that based on large sample approximations.

KAPPA COEFFICIENT

The kappa statistic can be interpreted as a measure of agreement that exists beyond the amount expected by chance alone (Cohen, 1960). Consider a square k x k probability matrix **M** with elements p_{ij}, where the rows (indexed by i) correspond to the observations of rater 1 and the columns (indexed by j) to rater 2. Let

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

be the proportion of subjects in the ith row, and let

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

be the proportion of subjects in the jth column. Let w_{ij}, $0 \le w_{ij} \le 1$, be the weight assigned to cell i,j. Then the observed agreement between the two raters is

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

and the amount of agreement expected by chance alone is

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

The weighted kappa coefficient is defined as

$$\hat{\kappa}_w = (p_o - p_c)/(1 - p_c).$$

Note that the simple kappa coefficient is a special case of $\hat{\kappa}_w$, with $w_{ij}=1$ for i=j and $w_{ij}=0$ for $i \neq j$.

Values of kappa and weighted kappa generally range from 0 to 1, although negative values are possible. A value of 1 indicates perfect agreement, while a value of 0 indicates no additional agreement beyond what is expected by chance alone. A negative value of kappa indicates agreement which is less than expected by chance alone.

THE BOOTSTRAP

The bootstrap (Efron, 1982) is a resampling technique for estimating the precision of a parameter estimate. This method provides an alternative to large sample techniques when asymptotic properties are not met or when the standard error of the estimate has complicated mathematical characteristics.

Consider an independent and identically distributed sample of size n from an unknown probability distribution F, and let $\kappa = \kappa(F)$ be the unknown population kappa coefficient. $\hat{\kappa}$, the parameter estimate of κ, can be calculated as defined above using the observed data $X_1=x_1, X_2=x_2, ..., X_n=x_n$. Efron (1982) describes two methods to create a nonparametric confidence interval about this estimate using bootstrap methodology: the percentile method and the bias-corrected percentile method. Both methods require the same initial steps:

1) Construct $\hat{F}$, the empirical distribution function of the observed data. $\hat{F}$ places probability 1/n on each observed data point $x_1, x_2, ..., x_n$.
2) Draw a bootstrap sample $X_1^*, X_2^*, ..., X_n^*$ of size n with replacement from $\hat{F}$. Then calculate $\hat{\kappa}^* = \hat{\kappa}(X_1^*, X_2^*, ..., X_n^*)$.
3) Repeat step (2) a large number of times, say 1000, and then rank the values $\hat{\kappa}^*$.

The percentile method is the least complicated of the two methods for calculating a confidence interval. For a $1-2\alpha$ interval, after ranking the bootstrapped kappa coefficients, simply take the α th percentile as the lower confidence limit and the $(1-\alpha)$th percentile as the upper confidence limit. The bias-corrected percentile method is slightly more complicated and takes into account the possible skewness of the sampling distribution of kappa. Briefly, this method entails determining the proportion of bootstrapped estimates that fall below the observed estimate and incorporating this asymmetry into the selection of percentiles used to determine the confidence limits. More detail can be found in Efron (1982).

Kappa by definition is bounded by 1, meaning that its sampling distribution becomes progressively skewed to the left as kappa approaches 1. The asymptotic confidence interval does not take this skewness into account and can produce upper confidence limits that exceed 1. A large sample size can partially alleviate this problem, but the required sample size grows exponentially as kappa increases. Since the bootstrapped intervals are produced by percentiles of samples based on the observed data, an upper confidence limit exceeding 1 is not possible, indicating that bootstrapped intervals may be more appropriate when kappa is large and sample size is small.

SAS MACRO

The macro BKAPPA (appendix) calculates bootstrapped confidence intervals for a kappa statistic and compares them with the asymptotic confidence limits presented in PROC FREQ. Confidence limits for both the percentile method and the bias-corrected percentile method are calculated. Keyword parameters are required to specify the data set, the two variables (raters) used to generate a kappa statistic, the alpha level used to create the confidence interval, the upper and lower confidence limits, the type of kappa statistic, and the number of bootstrapped replications. The type of kappa statistic can be one of two values: KAPPA for a simple kappa coefficient and WTKAP for a weighted kappa coefficient. If the latter is specified, a form similar to that of Cicchetti and Allison (1971) is used to create weights. The macro calculates kappa and weighted kappa coefficients for the observed data and bootstrap samples using the AGREE option in PROC FREQ. A requirement of this calculation is that the number of categories for rater 1 equals the number of categories for rater 2. It is possible that a bootstrapped sample may not include values from each rater for each category, especially if cells are sparse in the observed data set. If kappa cannot be estimated from a bootstrapped sample, then that sample is excluded from the confidence interval calculation.

EXAMPLE

Table 1 contains SAS output for a hypothetical data set called RATE comparing the results of two raters, RATER1 and RATER2. The estimated kappa coefficient and large sample 95% confidence interval are also included, and are produced using the following SAS code:

```
proc freq data=rate;
   tables rater2*rater1 / agree alpha=.05;
run;
```

```
 Table 1:  Kappa Statistics and Large Sample
      Standard Error for Data Set Rate

RATER2     RATER1

 Frequency|
 Percent  |
 Row Pct  |
 Col Pct  |       1|       2|       3|  Total
 ---------+--------+--------+--------+
        1 |      9 |      0 |      1 |     10
          |  37.50 |   0.00 |   4.17 |  41.67
          |  90.00 |   0.00 |  10.00 |
          |  90.00 |   0.00 |  10.00 |
 ---------+--------+--------+--------+
        2 |      1 |      3 |      1 |      5
          |   4.17 |  12.50 |   4.17 |  20.83
          |  20.00 |  60.00 |  20.00 |
          |  10.00 |  75.00 |  10.00 |
 ---------+--------+--------+--------+
        3 |      0 |      1 |      8 |      9
          |   0.00 |   4.17 |  33.33 |  37.50
          |   0.00 |  11.11 |  88.89 |
          |   0.00 |  25.00 |  80.00 |
 ---------+--------+--------+--------+
 Total          10        4       10       24
             41.67    16.67    41.67   100.00


 STATISTICS FOR TABLE OF RATER2 BY RATER1

            Test of Symmetry
            ----------------
Statistic = 2.000   DF = 3   Prob = 0.572


             Kappa Coefficients

Statistic      Value  ASE    95% Bounds
---------------------------------------
Simple Kappa   0.738 0.117   0.509 0.967
Weighted Kappa 0.784 0.106   0.576 0.992

Sample Size = 24
```

The bootstrapped 95% confidence intervals are obtained using the following macro call:

```
%bkappa(ds=rate,xvar=rater1,
   yvar=rater2,alpha=.05,p1=2.5,p2=97.5,
   kappa=WTKAP,sim=1000)
```

The output produced by macro BKAPPA appears in Table 2. Notice that the observed kappa coefficient is relatively large, indicating a possible skewed distribution. The large sample confidence interval does not account for this and thus is shifted to the right. The percentile method and the bias-corrected percentile method produce results that may more accurately reflect the skewness of the sampling distribution.

```
 Table 2:  Output Produced by BKAPPA Macro

Data Set:                       rate
Type of Kappa Coefficient:      wtkap
Lower CI Percentile:            2.5
Upper CI Percentile:            97.5
Rater 1:                        rater1
Rater 2:                        rater2


Estimated Kappa Coefficient:    0.78
Asymptotic Standard Error:      0.11


          CONFIDENCE INTERVALS
Large Sample:              (0.58 , 0.99)
Percentile Method:         (0.53 , 0.96)
Bias-Corrected Method:     (0.52 , 0.96)
```

CONCLUSION

The kappa statistic has become a standard tool for measuring the amount of agreement between two sources, but its asymptotic standard error should be used with caution when sample size is inadequate. Bootstrap resampling provides a viable alternative to construction of confidence limits when large sample approximations cannot be used or calculation of standard errors is analytically complex. The growing accessibility of high speed computers further encourages the use of bootstrap resampling methods for these situations.

REFERENCES

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.

Cohen, J. (1968) “Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit,” *Psychological Bulletin*, 70, 213-220.

Efron, B. (1982), *The jackknife, the bootstrap, and other resampling plans*, CBMS 38, SIAM-NSF.

Fleiss, J.L., Cohen, J., and Everitt, B.S. (1969), “Large sample standard errors of kappa and weighted kappa,” *Psychological Bulletin*, 72, 323-327.

Fleiss, J.L. and Cicchetti, D.V. (1978), “Inference about weighted kappa in the non-null case,” *Applied Psychological Measurement*, 2, 113-117.

Fleiss, J.L. (1981), *Statistical methods for rates and proportions, second edition*, New York: John Wiley & Sons, Inc.

SAS Institute Inc., *SAS/STAT™ software: changes and enhancements through release 6.11*, Cary, NC: SAS Institute, Inc., 1996.

CONTACT INFORMATION:

Robert A. Vierkant, MAS
Marshfield Medical Research Foundation
1000 N. Oak Avenue
Marshfield, WI 54449
(715) 389-3536
Email: vierkanr@mfldclin.edu

APPENDIX: SAS MACRO BKAPPA

```
***************************************;
**                                   **;
**         SAS Macro BKAPPA          **;
**     parameters are as follows     **;
**                                   **;
** 1) ds=data set (default=last)     **;
** 2) xvar=rater 1 (horizontal axis) **;
** 3) yvar=rater 2 (vertical axis)   **;
** 4) alpha=alpha level of interval  **;
**       (e.g., .05)                 **;
** 5) p1=lower confidence limit      **;
**       (e.g., 2.5)                 **;
** 6) p2=upper confidence limit      **;
**       (e.g., 97.5)                **;
** 7) kappa=type of kappa statistic  **;
**         options are kappa, wtkap  **;
** 8) sim=number of bootstrapped     **;
**         samples drawn             **;
**                                   **;
***************************************;

%macro bkappa(ds=_last_,xvar=,yvar=,
              alpha=.05,p1=2.5,p2=97.5,
              kappa=kappa,sim=1000);
****initialize the bootstrap data set;
data boot2; set _null_; run;

****macro variable of observations
    per bootstrapped sample;
data &ds; set &ds nobs=nobs;
   id=_n_;
   call symput('num',left(trim(nobs)));
run;

****get the kappa statistics on the
    original data;
proc freq data=&ds noprint;
   tables &yvar*&xvar
   / agree alpha=&alpha;
   output out=boot1 &kappa;
run;

****macro variable for observed kappa;
data boot1; set boot1;
  call symput('k',left(trim(_&kappa._)));
run;
```

```
****generate bootstrap samples;
data samp (drop=i);
   do sim=1 to &sim;
      do i=1 to &num;
         id=int(&num*ranuni(-1)+1);
         set &ds point=id;
         if _error_ then abort;
         output;
      end;
   end;
   stop;
run;

****generate kappa statistics for
    bootstrap samples;
proc freq data=samp noprint;
   by sim;
   tables &yvar*&xvar / agree;
   output out=boot2 &kappa;
run;

****generate macro variables p3 & p4
    for bias-corrected percentile method;
data boot3;
   set boot2 end=last;
   k=symget('k'); p1=symget('p1');
   if . lt _&kappa._ lt k then n+1;
   if _&kappa._ gt . then d+1;
   if last then do;
     zalpha=probit(1-p1/100);
     z0=probit(n/d);
     p3=100*probnorm(2*z0-zalpha);
     p4=100*probnorm(2*z0+zalpha);
     call symput('p3',trim(left(p3)));
     call symput('p4',trim(left(p4)));
     output;
   end;
run;

****delete any bootstrap samples that
    did not produce legitimate kappa;
data boot2; set boot2;
   if _&kappa._=. then delete;
run;

****get the p1th, p2th, p3th, and p4th
    percentiles of bootstrap samples;
proc sort data=boot2; by _&kappa._; run;
data results (keep=boot1--boot4);
  set boot2 nobs=last;
  by _&kappa._;
  id=_n_;
  if id=1 then do;
    boot1=0; boot2=0; boot3=0; boot4=0;
  end;
  if id/last le &p1/100 and (id+1)/last
    ge &p1/100 then boot1=_&kappa._;
  if id/last le &p2/100 and (id+1)/last
    ge &p2/100 then boot2=_&kappa._;
  if id/last le &p3/100 and (id+1)/last
    ge &p3/100 then boot3=_&kappa._;
  if id/last le &p4/100 and (id+1)/last
    ge &p4/100 then boot4=_&kappa._;

  if id=last then output;
  retain boot1 boot2 boot3 boot4;
run;

****merge percentiles in with
    original kappa;
data boot; merge boot1 results; run;

****print out results;
options nodate;
data _null_;
   set boot;
   file print;
   put @1 ///
       @5 'Data Set:'
       @55 "&ds" /
       @5 'Type of Kappa Coefficient:'
       @55 "&kappa" /
       @5 'Lower CI Percentile:'
       @55 "&p1" /
       @5 'Upper CI Percentile:'
       @55 "&p2" /
       @5 'Rater 1:'
       @55 "&xvar" /
       @5 'Rater 2:'
       @55 "&yvar" ///
       @5 'Estimated Kappa Coefficient:'
       @55 _&kappa._ 6.2 /
       @5 'Asymptotic Standard Error:'
       @55 e_&kappa 6.2 ///
       @25 'CONFIDENCE INTERVALS' /
       @5 'Large Sample:'
       @51 '(' l_&kappa 6.2
       @58 ',' u_&kappa 6.2 ')' /
       @5 'Percentile Method:'
       @51 '(' boot1 6.2
       @58 ',' boot2 6.2 ')' /
       @5 'Bias Corrected Percentile Method:'
       @51 '(' boot3 6.2
       @58 ',' boot4 6.2 ')';
run;
%mend bkappa;
```