Kappa Statistics for Multiple Raters Using Categorical Classifications

Annette M. Green, Westat, Inc., Research Triangle Park, N.C.

ABSTRACT

In order to assess the reliability of a given characterization of a subject it is often necessary to obtain multiple readings, usually but not always from different individuals or raters. The degree of agreement among the various raters gives some indication as to the consistency of the values. If agreement is high, we feel more confident the ratings reflect the actual circumstance. If agreement among the raters is low, we are less confident in the results. While several methods are available for measuring agreement when there are only two raters, this paper concentrates on presenting a generalized implementation of the Fleiss (1981) technique. This method can be utilized even in situations where there are more than two raters and/or categories. A review of the statistical theory behind the intraclass correlation coefficients and kappa statistics obtained when looking at the above situations is presented. SAS® code is provided which utilizes basic SAS procedures.

INTRODUCTION

Often we are faced with determining the measurement of interrater agreement when the ratings are on a categorical scale. When the number of raters is equal to two, this is easily accomplished by using SAS PROC CORR to get an estimate of the correlation coefficient. SAS 6.10 PROC FREQ with the AGREE option also provides an easy way to obtain the kappa statistic when there are only two raters. Fleiss describes a technique for obtaining interrater agreement when the number of raters is greater than or equal to two. This paper concentrates on the ability to obtain a measure of agreement when the number of raters is greater than two. It also concentrates on the technique necessary when the number of categories into which the ratings can fall is greater than two.

BACKGROUND

The data used in this paper to demonstrate the technique of calculating interrater agreement was compiled from a gastric graft versus host disease study. (Washington, *et al*, 1996, submitted). This research compares the results of three pathologists' diagnoses of 51 different gastric biopsies. Each pathologist reviewed the gastric biopsies in a blinded fashion. The specific data used in the example in this paper involves degree of agreement on density of the inflammatory infiltrate in the lamina propria. The categories of degree of inflammation range from zero to three.

METHODS

Since the kappa statistic ($\hat{k}$) was first proposed by Cohen (1960), variants have been proposed by others, including Scott (1955), Maxwell and Pilliner (1968), and Bangdiwala (1987). SAS code has also been presented by Gaccione (1993) to compute the kappa statistic. The various discussions describe the kappa statistic as being very similar, if not equivalent, to the *intraclass correlation coefficient* (Ebel, 1951).

One of the most important features of the kappa statistic is that it is a measure of agreement which naturally controls for chance. Since its development, there has been much discussion on the degree of agreement due to chance alone. According to Fleiss, there is a natural means of correcting for chance using an *indices of agreement*. Kappa is based on these indices. If there is complete agreement, $\hat{k} = 1$. If the observed agreement is greater than or equal to chance agreement, $\hat{k} \geq 0$, and if the observed agreement is less than or equal to chance agreement, $\hat{k} \leq 0$.

When a composite measure of agreement across all categories is desired, an overall value of kappa may be used. This is defined as a weighted average of individual kappa values and is the basis for the procedure described in this paper.

Before demonstrating the SAS code necessary to create the desired kappa statistic, a review of the theory is warranted. Consider a sample of (n) subjects which have been rated independently by two or more different raters (m). According to Fleiss, the raters responsible for rating one subject do not have to be the same as those responsible for rating another. Let (m_i) be the number of ratings on the ith subject and (k) be the number of categories into which classifications can be made. The following review of Fleiss' theory and the subsequent SAS code concentrate on the case of greater than two raters and greater than two categories into which the ratings can be classified. If the ratings and/or categories are equal to two, the following formulae can be simplified to accommodate those circumstances. A discussion of the simplified versions of these formulae will follow the more complicated case. As mentioned above, the AGREE option in SAS 6.10 PROC FREQ makes calculation of the kappa statistic readily available when the number of raters is equal to two.

Values for an individual kappa per category ($\hat{k}_j$) and an overall kappa ($\hat{\bar{k}}$) are necessary in order to test the hypothesis that the underlying value of kappa is zero (i.e., the ratings are independent). The following theory is appropriate when the number of ratings per subject is constant and equal to m.

Define x_{ij} as the number of ratings on subject $i(i = 1, \ldots, n)$ into category $j(j = 1, \ldots, k)$ where

$$\sum_{j=1}^{k} x_{ij} = m \qquad (1)$$

for all i.

Let the mean number of ratings per subject be $\bar{m}$. Note, if the number of raters are equal for each subject $\bar{m}$ will equal m.

$$\bar{m} = \frac{\sum_{i=1}^{n} m_i}{n} \qquad (2)$$

Let $\bar{p}_j$ denote the overall proportion of ratings (observed agreement) in category *j* and $\hat{k}_j$ the value of kappa for category *j*, *j* = 1, . . . , *k*.

$$\bar{p}_j = \frac{\sum_{i=1}^{n} x_i}{n\bar{m}} \qquad (3)$$

The value of $\hat{k}_j$ is then

$$\hat{k}_j = 1 - \frac{\sum_{i=1}^{n} x_{ij}(m - x_{ij})}{nm(m-1)\,\bar{p}_j\bar{q}_j}, \qquad (4)$$

Thus this formula is a measure of interrater agreement per category, where $\bar{q}_j = 1 - \bar{p}_j$.

As discussed earlier, an overall value of kappa may be defined as a weighted average of the individual kappa values. Landis and Koch (1977) described the weighted average below where the weights are the denominators of the individual kappas.

$$\hat{\hat{k}} = \frac{\sum_{j=1}^{k} \bar{p}_j\bar{q}_j\hat{k}_j}{\sum_{j=1}^{k} \bar{p}_j\bar{q}_j} \qquad (5)$$

which is equivalent to

$$\hat{\hat{k}} = 1 - \frac{nm^2 - \sum_{i=1}^{n}\sum_{j=1}^{k} x_{ij}{}^2}{nm(m-1)\sum_{j=1}^{k} \bar{p}_j\bar{q}_j}. \qquad (6)$$

Fleiss (1971) showed algebraically equivalent versions of these formulae which demonstrated explicitly how they represent chance-corrected measures of agreement. The statistical methods described in this paper for controlling for error are applicable only when the rates of misclassification are known from an external source or are estimable by applying a standard classification procedure to a subset of the group.

EXAMPLE

Although the above theory looks complicated the formulae presented can be easily generated using SAS Base or SAS Macro. Due to the simplicity of formula (6) versus (5) the following SAS code will concentrate on the more basic formula. Detailed explanation of the basic code is given below.

In our example there are 51 subjects. Three different pathologists (raters) classified gastric biopsies from each patient into one of four categories. Following the nomenclature used in the above theory, assignment is as follows: n = 51, m = 3, and k = 4.

The sum of each category into which raters can classify the subject are represented by CAT_X. In our example there are four separate categories, 0, 1, 2, or 3; categories names are therefore CAT_0, CAT_1, CAT_2, and CAT_3 respectively.

The data step begins by entering initial rater readings and the summation of each category total for each subject. In the following example these variables are referred to as RATER_1, RATER_2, RATER_3, CAT_0, CAT_1, CAT_2, and CAT_3. The values for sample size (n), number of raters (m), and number of categories (k) are then initialized. It is then necessary to create variables which will be needed in the kappa calculations. The first of these variables is x_{ij}^2 which is referred to as sq_x in the SAS code. The other variable is $x_{ij}(m\text{-}x_{ij})$, this is part of the numerator necessary to calculate the $\hat{k}_j$ statistic. The number of variables necessary for this variable ranges from $j = (1, \ldots, k)$. In this example where k equals four, the variables are named A_0, A_1, A_2, and A_3 respectively.

```
data one;
  [Enter initial readings and sum-
mation variables for each category per
subject.]

  [Assign values for n, m, and k]

*
  SQ_X is necessary in order to derive
the summation from j = 1 to k of the
individual  squares (x_ij)
;

sq_x = ((cat_0**2) + (cat_1**2) +
        (cat_2**2) + (cat_3**2));
```

Summation of many of the variables created in the data step are necessary to continue calculation of the kappa statistic. PROC MEANS will provide these summations quickly and place the necessary data in an output file which will be referred to as Results in this example.

```
proc means data=one;
  var cat_0 cat_1 cat_2 cat_3 sq_x;
  output out=results
        sum(cat_0) = sum_cat0
        sum(cat_1) = sum_cat1
        sum(cat_2) = sum_cat2
        sum(cat_3) = sum_cat3
        sum(sq_x)  = sum_sq_x
   ;
*
The separate summation variables for
each CAT_X variable are necessary in
order to calculate the proportions of
ratings in each category.

SUM(SQ_X) is the summation from
i = 1 to n of the sum of j = 1 to k of x_ij^2
;
run;
```

It is now necessary to create a second data set which will use the summation values created by PROC MEANS along with other variables which are necessary to complete the calculations. The n and m variables must be reentered here since they are not included in the Results file.

```
*
Refer to formula 3 in the Methods
Section for calculation of proportion
variables.
;

pbar_dem = (n * m);
*Denominator for p_bar variables;

  p_bar0 = sum_cat0/pbar_dem;
  p_bar1 = sum_cat1/pbar_dem;
  p_bar2 = sum_cat2/pbar_dem;
  p_bar3 = sum_cat3/pbar_dem;

  q_bar0 = 1 - p_bar0;
  q_bar1 = 1 - p_bar1;
  q_bar2 = 1 - p_bar2;
  q_bar3 = 1 - p_bar3;


  pq_cat0 = p_bar0 * q_bar0;
  pq_cat1 = p_bar1 * q_bar1;
  pq_cat2 = p_bar2 * q_bar2;
  pq_cat3 = p_bar3 * q_bar3;
  sum_pq = sum(of pq_cat0-pq_cat3);


*
  Calculate overall kappa referred to
as KAPPA.  Refer to formula 6.
;

kappa = 1 - ((n * (m**2) - sum_sq_x)/
          (n * m * (m - 1) * sum_pq));

proc print;
  var   kappa;
run;
```

RESULTS

The general consensus is that kappa values greater than 0.75 are considered to have a high degree of agreement beyond chance. Values below 0.40 have a low degree of agreement and values between 0.40 and 0.75 represent a fair to good level of agreement beyond chance alone.

The results of our particular study calculate the overall kappa (KAPPA) to be equal to 0.218. Therefore suggesting that the agreement among the three raters in detecting degree of inflammatory infiltrate in the lamina propria is low. Although the code has not been provided other variables can be calculated using similar coding for formula 5. Following the necessary steps to calculate this formula will produce the individual kappas for each category as well as the overall kappa.

In cases where the number of ratings per subject is equal Fleiss, Nee, and Landis (1979) derived and confirmed formulae for the approximate standard errors of $\hat{\bar{k}}$ and $\hat{k}_j$, each appropriate for testing the hypothesis that the underlying value of kappa is zero. The formulae are as follows:

$$s.e._0(\hat{\bar{k}}) =$$

$$\frac{\sqrt{2}}{\sum_{j=1}^{k} \bar{p}_j \bar{q}_j \sqrt{nm(m-1)}} X \sqrt{\left(\sum_{j=1}^{k} \bar{p}_j \bar{q}_j\right)^2 - \sum_{j=1}^{k} \bar{p}_j \bar{q}_j (\bar{q}_j - \bar{p}_j)} \qquad (7)$$

and

$$s.e._0(\hat{k}_j) = \sqrt{\frac{2}{nm(m-1)}} \qquad (8)$$

It is important to note that $s.e._0(\hat{k}_j)$ is independent of $\bar{p}_j$ and $\bar{q}_j$. It is also important to remember the above formulae are valid when the number of ratings per subject are the same. To this author's knowledge, the standard error of $\hat{\bar{k}}$ has not yet been derived when the numbers of ratings per subject vary.

CONCLUSION

With the advent of the AGREE option in SAS 6.10 PROC FREQ we can now calculate a kappa between two raters with ease. When the number of raters are greater than two another method must be employed. The discussion presented above simply transforms an accepted theory for the calculation of the kappa statistic into simple SAS code. An understanding of the underlying theory and a basic knowledge of SAS should enable a user easy access to this procedure.

If version 6.10 or higher is not available to the user, the given formulae can be simplified to conform to the case where the number of raters are equal to two.

REFERENCES

Bangdiwala, S.I., Bryan, H.E. (1987), "Using SAS Software Graphical Procedures for the Observer Agreement Chart," *Proceedings of the Twelfth Annual SAS Users Group International Conference*, 12, 1083-1088.

Cohen, J. (1960), A Coefficient of Agreement for Nominal Scales. *Educ. Psychol. Meas*., 20, 37-46.

Ebel, R.L. (1951), Estimation of the Reliability of Ratings. *Psychometrika*, 16, 407-424.

Fleiss, J.L. (1971), Measuring Nominal Scale Agreement Among Many Raters. *Psychol. Bull*., 76, 378-382.

Fleiss, J.L. (1981), The Measurement of Interrater Agreement, *Statistical Methods for Rates and Proportions, Second Edition*, John Wiley & Sons, Inc., New York 212-304.

Fleiss, J.L., Nee, J.C.M., and Landis, J.R. (1979), The Large Sample Variance of Kappa in the Case of Different Sets of Raters. *Psychol. Bull*., 86, 974-977.

Gaccione, P. (1993), "Data Step Kappa Computations," *Proceedings of the Eighteenth Annual SAS Users Group International Conference*, 18, 461-466.

Landis, J.R. and Koch, G.G. (1977), A One-Way Components of Variance Model for Categorical Data. *Biometrics*, 33, 671-679.

Maxwell, A.E. and Pilliner, A.E.G. (1968), Deriving Coefficients of Reliability and Agreement for Ratings. *Br. J. Math. Stat. Psychol*., 21, 105-116.

Scott, W.A. (1955), Reliability of Content Analysis: The Case of Nominal Scale Coding. *Public Opinion Quart*., 19, 321-325.

Washington, K., Bentley, R.C., Green, A.M., Olson, J., Treem, W.R., Krigman, H.R., (1996), Gastric Graft-Versus-Host Disease: A Blinded Histologic Study, *American Journal of Surgical Pathology*, submitted.

ACKNOWLEDGMENTS

I would like to thank Drs. Kay Washington, Rex Bentley, and Hannah Krigman , Dept. of Pathology at Duke University Medical Center in Durham, N.C. for the use of their data.

For further information contact Annette Green, green3@niehs.nih.gov.