Series: Statistical Methods For Inter-Rater Reliability Assessment, No. **3**, October 2002

Computing Inter-Rater Reliability With the SAS System

Kilem Gwet, Ph.D.
Sr. Statistical Consultant, STATAXIS Consulting
kilem62@yahoo.com

Abstract.

The SAS system V.8 implements the computation of unweighted and weighted kappa statistics as an option in the FREQ procedure. A major limitation of this implementation is that the kappa statistic can only be evaluated when the number of raters is limited to 2. Extensions to the case of multiple raters due to Fleiss (1971) have not been implemented in the SAS system. A SAS macro called MAGREE.SAS, that can handle the case of multiple raters is available at the SAS Institute's web site (check it at http://ftp.sas.com/techsup/download/stat/magree.sas).

In this article, we discuss about the use of the SAS system to compute Kappa statistics in general. We will also present our SAS macro called `INTER_RATER.MAC` that can handle multiple raters and can compute AC1 and Kappa statistics (overall and for each category) as well as the associated standard errors. In the `INTER_RATER.MAC` SAS macro, the AC1 standard error is calculated both conditionally on the sample od raters as well as unconditionally. The unconditional standard error has the particular feature of taking into account the additional variability that is due to the sampling of raters.

1. **Introduction**

There is a confusion about the definition of the Kappa statistic. Cohen (1960) first proposed the Kappa statistic as a way to evaluate the extent of agreement between raters. His article came five years after Scott (1955) suggested the PI-statistic (or π-statistic) as a measure of the inter-rater reliability for two raters. Fleiss (1971) extended inter-rater reliability assessment to the case of multiple raters with multiple response categories and referred to it as the generalized Kappa statistic. Here is where the confusion originates. Fleiss' proposal is actually a generalization of Scott's π-statistic rather than that of Cohen's Kappa. Our claim is justified by the fact that Fleiss' generalized Kappa statistic reduces to Scott's π-statistic when the number of raters is 2. There is a possibility that Kappa being a well-known statistic, it is the name Fleiss had in ming as he was studying a way to assess agreement between multiple raters.

The Kappa statistic that is implemented in the SAS system for two raters is the original proposal by Cohen (1960). The generalized Kappa statistic that is implemented in the MAGREE.SAS macro is the generalized version of Fleiss (1971). If the `MAGREE.SAS` macro is used with two-rater reliability data, it will not yield the same result as the AGREE option of the SAS FREQ procedure. This is due to the fact that Fleiss' generalized Kappa statistic reduces to Scott's π-statistic rather than to Cohen's Kappa. For the sake of clarity, I believe that the generalized PI (or $\pi-$) statistic of Fleiss (1971) should be referred to as "Generalized π-statistic". Gwet

(2001) discusses in chapter 5, what a natural extension of Cohen's Kappa statistic to multiple raters would look like.

2. Unweighted κ/π Statistic

In a typical two-rater and q response category reliability study involving raters A and B and categories 1 to q, the data will be represented in a $q \times q$ contingency table as shown in Table 1. The number n_{21} for instance, indicates that there are n_{21} subjects that raters A and B have classified into categories 2 and 1 respectively. The total number of subjects rated is denoted as n.

We assume throughout this paper that there is no missing rating. That is, each rater is assumed to have rated all n subjects that participated to the reliability experiment. Consequently, marginal totals are expected to sum to the subject sample size n. Although the problem of missing ratings may occur in some reliability studies, it can only be dealt with using special techniques that are beyond the scope of this paper.

Table 1: Distribution of n subjects by rater and response category

Rater A	Rater B				Total
	1	2	$\cdots$	q	
1	n_{11}	n_{12}	$\cdots$	n_{1q}	n_{1+}
2	n_{21}	n_{22}	$\cdots$	n_{2q}	n_{2+}
$\vdots$			$\cdots$		$\vdots$
q	n_{q1}	n_{q2}	$\cdots$	n_{qq}	n_{q+}
Total	n_{+1}	n_{+2}	$\cdots$	n_{+q}	n

Let P_a be the overall agreement probability. It represents the probability that raters A and B classify a randomly selected subject into the same category and is given by:

$$p_a = \sum_{k=1}^{q} p_{kk} \text{ where } p_{kk} = n_{kk}/n. \tag{1}$$

Scott's π-statistic is given by:

$$\mathcal{PI} = \frac{p_a - p_e(\pi)}{1 - p_e(\pi)}, \tag{2}$$

where $p_e(\pi)$ measures the likelihood of agreement by chance. Using the information contained in Table 1, $p_e(\pi)$ can be expressed as follows:

$$p_e(\pi) = \sum_{k=1}^{q} p_k^2, \text{ where } p_k = (p_{+k} + p_{k+})/2, \tag{3}$$

and $p_{+k} = n_{+k}/n$, and $p_{k+} = n_{k+}/n$.

Cohen's κ-statistic is given by:

$$\mathcal{KA} = \frac{p_a - p_e(\kappa)}{1 - p_e(\kappa)}, \tag{4}$$

where $p_e(\kappa)$ provides Cohen's measure of the likelihood of agreement by chance. Using the information contained in Table 1, $p_e(\kappa)$ can be expressed as follows:

$$p_e(\kappa) = \sum_{k=1}^{q} p_{+k}p_{k+}. \tag{5}$$

Cohen's Kappa statistic uses rater-specific classification probabilities p_{k+} and p_{+k} to compute the likelihood of agreement by chance, while Scott's approach is based on the overall classification probabilities p_k.

2. Weighted κ/π Statistic

In some applications, researchers may want to treat possible disagreements between raters differently. This will usually be the case when the response categories of Table 1 are ordered. This problem is resolved by using a generalized version of the Kappa statistic referred to as the weighted Kappa introduced by Cohen (1968). The general form of the weighted Kappa statistic, denoted by $\mathcal{K}_a'$, is given by:

$$\mathcal{K}_a' = \frac{p_a' - p_e'(\kappa)}{1 - p_e'(\kappa)}, \tag{6}$$

where the overall agreement probability p_a' and chance-agreement probability $p_e'(\kappa)$ are respectively defined as follows:

$$p_a' = \sum_{k=1}^{q}\sum_{l=1}^{q} w_{kl}p_{kl}, \tag{7}$$

$$p'_e(\kappa) = \sum_{k=1}^{q}\sum_{l=1}^{q} w_{kl}p_{k+}p_{+l}. \tag{8}$$

In equation(7), $p_{kl} = n_{kl}/n$ represents the proportion of subjects classified into the (k,l)-th cell. In equation (8), $p_{k+} = n_{k+}/n$ and $p_{+l} = n_{+l}/n$ represent respectively the proportions of subjects that raters A and B classified into categories k and l.

Readers interested in the standard errors of Kappa and weighted Kappa should read the paper by Fleiss et al. (1969). The formulas implemented in the SAS system are all presented in that paper.

3. Computing κ/π Statistic with the FREQ Procedure of SAS

The SAS system as of version 8 implements the computation of Cohen's Kappa and weighted Kappa statistics (when the number of raters is limited to 2 but does not implement that of Scott's $\mathcal{PI}$ statistic. In order to compute the kappa statistics and associated statistical tests and precision measures, it is necessary to prepare a SAS data set as shown in Table 2. The Input file would have three variables containing the subject identification, and the scores of both raters A and B. Although the "Subject" variable is not mandatory, it is useful for relating the ratings to specific subjects. In Table 2, Category(A,2) for instance represents the category (any number from 1 to q if the categories are labelled as 1,2,...,q) into which rater A has classified subject 2.

Table 2. Input file for computing the κ statistic With SAS

Subject	Rater_A	Rater_B
1	Category$(A,1)$	Category$(B,1)$
2	Category$(A,2)$	Category$(B,2)$
$\vdots$	$\vdots$	$\vdots$
n	Category(A,n)	Category(B,n)

Figure 1 provides an example of a SAS data step, which describes a data set (ClassData.sas7bdat) showing how 2 raters $R1$ and $R2$ classified 10 subjects into one of two possible categories labelled as $+$ and $-$. In this case, the variables $R1$ and $R2$ are of alphabetic type, but could be numeric.

```
DATA ClassData;
     INPUT Subject$ R1$ R2$;
     DATALINES;
          1 - +
          2 - -
          3 + -
          4 - -
          5 - -
          6 - +
          7 - -
          8 + +
          9 - -
         10 - -
         ;
```

Figure 1. Data step for defining input data set.

The FREQ procedure is the official gateway for obtaining inter-rater reliability estimates and associated precision measures with the SAS system. The SAS user should note that this procedure can only be used when the number of raters is limited to 2. If the number of raters is greater than 2, then special SAS programs must be used (more on this in subsequent sections).

3.1) Commands for obtaining kappa

Using the classification data reported in figure 1, the extent of agreement between raters $R1$ and $R2$ as measured by the kappa statistic (see equation 4), is obtained by submitting the statements shown in example 1:

Example 1: Kappa Coefficient without P-values

```
      PROC FREQ DATA = ClassData;
           TABLES R1*R2/AGREE;
           RUN;
```

The outcome of this program is described in figure 2. The "AGREE" option in the TABLES statement can be replaced with the "KAPPA" option and the outcome will be the same. It appears from figure 2 that SAS only calculated the simple Kappa statistic and not the weighted Kappa. This may be surprising as the AGREE or KAPPA options are supposed to produce simple and weighted kappa statistics, with the Cicchetti-Allison weights by default (more on weighted kappa later in this section). The weighted kappa statistic is not computed in this case because it is always identical to simple kappa when the number of response categories is limited to 2 for any of the weights offered by SAS.

- The first part of the output is a frequency table showing the distribution of subjects by rater and response category.
- The second part of the output provides the results of the marginal homogeneity testing. Since the frequency table is a **2** × **2** table, the McNemar's test statistic is used. For bigger tables, marginal homogeneity is tested using the Bowker's test statistic. Both statistics are equivalent on **2** × **2** tables. The most important statistic to look at in this table is the p-value. That is `Pr>S = 0.5637`. When this number is smaller than 0.05, one may conclude that the marginals are not homogeneous. That is, there is not a strong enough evidence to support the fact that the raters may have the same rating propensities. In this example the hypothesis of marginal homogeneity is not rejected. We are not convinced about the usefulness of this hypothesis test. In fact, whether the marginals are homogeneous or not is irrelevant as far as the Kappa statistic is concerned. It does not affect neither its validity nor its precision. Some authors have linked this hypothesis test to the validity of the $\boldsymbol{\pi}$-statistic (see Zwick (1988)), but certainly not that of the Kappa statistic implemented in SAS.
- The last table provides the simple kappa estimate, the associated ASE (Asymptotic Standard Error) and the lower and upper bounds of the 95% confidence interval. The most important number in this table is the Kappa estimate, because it provides an estimate of the extent of agreement between raters. The next most important numbers are the two bounds of the confidence interval. The confidence interval contains the "true" kappa coefficient with a probability of 95%. However, the variance expression implemented in SAS is only valid when there does exist an agreement between raters beyond chance. This makes the interpretation of the confidence interval in figure 2 difficult as it contains **0** as a possible "true value".

```
R1            R2

Frequency|
Percent  |
Row Pct  |
Col Pct  |      +|      -|  Total
---------+-------+-------+
+        |      1|      1|      2
         |  10.00|  10.00|  20.00
         |  50.00|  50.00|
         |  33.33|  14.29|
---------+-------+-------+
-        |      2|      6|      8
         |  20.00|  60.00|  80.00
         |  25.00|  75.00|
         |  66.67|  85.71|
---------+-------+-------+
Total    |      3|      7|     10
            30.00   70.00  100.00

Statistics for Table of R1 by R2

     McNemar's Test
  -----------------------
  Statistic (S)  0.3333
  DF                  1
  Pr > S         0.5637

    Simple Kappa Coefficient
  --------------------------------
  Kappa                   0.2105
  ASE                     0.3282
  95% Lower Conf Limit   -0.4328
  95% Upper Conf Limit    0.8538

        Sample Size = 10
```

Figure 2. Output of SAS Procedure.

Let us consider a data set where categories are labelled so that they can ordered in a natural way. Figure 3 provides an example of such a data set. Two raters and three categories are defined in this data set. The three categories are "1", "2", and "3". They could have been labelled as "A", "B", and "C" and the result would have been the same.

```
DATA ClassData;
     INPUT Subject$ R1 R2;
     DATALINES;
         1 1 2
         2 1 1
         3 3 3
         4 2 2
         5 1 1
         6 2 2
         7 1 1
         8 2 2
         9 1 3
        10 1 1
         ;
```

Figure 3. Data step for defining an input data set with ordered categories.

After running the same FREQ procedure described in example *3.1*, one would obtain the results shown in figure **4**. The first table of figure 4 is a frequency table showing the distribution of subjects by raters and response category. The second table of figure **4** shows the results of marginal homogeneity testing. Since the number of categories in this example is 3 (i.e more than 2), the Bowker's test of symmetry is used. The P-value in this case (i.e. `Pr>S`) is evaluated at `0.5724`, which indicates that the "null" hypothesis of marginal homogeneity cannot be rejected. The data-based evidence does not suggest that the two raters `R1` and `R2` have different rating propensities. In the third table of figure 4, the simple and weighted Kappa statistics are shown as well as the associated standard errors (ASE) and 95%-confidence intervals. ASE stands for Asymptotic Standard Error, which represents a large-sample approximation of the standard error of the kappa statistic. The 95%-confidence interval provides a lower and an upper bounds that supposedly contain the "true" value of the kappa statistic with a 95% chance.

It is unclear how large the number of subjects should be for the ASE to provide a valid estimation of the standard error. Therefore, when using the FREQ procedure to compute the kappa statistic and its standard error, one should interpret the results with caution if the number of subjects is small.

3.2) Computing Weighted Kappa Statistics

The weighted kappa statistic is always computed whenever the `KAPPA` or `AGREE` option is specified in the `TABLES` statement of the `FREQ` procedure. SAS offers two types of weights for calculating weighted kappa. They are the Cicchetti-Allison (CA) and Fleiss-Cohen (FC) weights. The CA weights are defined for any two categories k and l by

$$w_{kl} = 1 - \frac{|C_k - C_l|}{C_q - C_1},$$

while the FC weights are given by:

$$w_{kl} = 1 - \frac{(C_k - C_l)^2}{(C_q - C_1)^2},$$

where C_k is a score attached to category k. The score types available in SAS are `TABLE`, `RANK`, `RIDIT` and `MODRIDIT`. "`TABLE`" scores represent either the numeric value of the category labels (if they are numeric), or the category numbers $(1, \cdots,$ q) if the labels are of character type. The "`RANK`" score for category k is given by:

$$C_k = \sum_{l<k} n_l + (n_k + 1)/2,$$

where n_l represents the number of subjects classified into category l. The "`RIDIT`" score for category k is given by:

$$C_k = \sum_{l<k} p_l + (p_k + 1/n)/2,$$

where p_l represents the percentage of subjects classified into category l. The modified ridit ("MODRIDIT") score for category k is defined as follows:

$$C_k = \frac{1}{n+1}\left\{\sum_{l<k} n_l + (n_k+1)/2\right\},$$

where n is the number of subjects in the sample.

The score to be used can be specified with `SCORES` option in the `TABLES` statement. The Cicchetti-Allison weights using the RIDIT score type can be specified as follows:

```
TABLES R1*R2/AGREE WT=CA SCORE=RIDIT.
```

Since the `CA` weights are used by defaults, the option `WT=CA` is unnecessary if the Cicchetti-Allison weights must be used. Fleiss-Cohen weights with `TABLE` scores can be obtained with the following statement:

```
TABLES R1*R2/AGREE WT=FC SCORE=TABLE.
```

3.3) Hypothesis Testing

In addition to computing the simple and weighted kappa statistics, researchers often want to known whether the obtained estimates are statistically significant. That is, are simple and weighted kappas statistically different from 0. The problem amounts to testing whether the observed extent of agreement between raters is not merely due to sampling variability and that there is no "real" agreement between the raters. The "null" hypotheses are "kappa=0" and "weighted kappa=0". These hypotheses can be tested by specifying the TEST option of the FREQ procedure as shown in example 2:

Example 2: Kappa Coefficient with P-values

```
PROC FREQ DATA = ClassData;
   TABLES R1*R2/AGREE;
   TEST AGREE;
   RUN;
```

After running this program on the ClassData file defined in figure 3, one will obtain in addition to the first two tables of figure 4, the four tables of figure 5. The contents of the first and third tables of figure 5 are identical to that of the third table of figure 4. However, the second and fourth tables of figure 5 provide the one-sided and two-sided P-values of the simple and weighted kappas respectively.

- ASE under H0 is the asymptotic standard error of the simple (or weighted) kappa when there is no agreement between the raters (i.e. simple kappa=0 or weighted kappa=0).
- Z represents the ratio of kappa estimate to its ASE under H0. This statistic follows the standard normal distribution under H0
- The one-sided P-value represents the probability that a variable that follows the standard normal distribution be greater than Z. Because the Z score follows the standard normal distribution if the inter-rater reliability equals 0, any P-value that is small (5% or smaller) indicates that the Z score is too large to be considered as normally-distributed variable with mean 0. In this case, the null hypothesis should be rejected.
- The two-sided P-value represents the probability that a variable that follows the standard normal distribution be greater the absolute value of the Z score.

If the user wants to test the statistical significance of simple kappa only and not that of weighted kappa, then the procedure in example 3 should be used.

Example 3: Kappa Coefficient P-values for Simple Kappa Only

```
PROC FREQ DATA = ClassData;
   TABLES R1*R2/AGREE;
   TEST KAPPA;
   RUN;
```

```
              Table of R1 by R2
R1               R2

Frequency|
Percent  |
Row Pct  |
Col Pct  |       1|       2|       3|  Total
---------+--------+--------+--------+
1        |       4|       1|       1|      6
         |   40.00|   10.00|   10.00|  60.00
         |   66.67|   16.67|   16.67|
         |  100.00|   25.00|   50.00|
---------+--------+--------+--------+
2        |       0|       3|       0|      3
         |    0.00|   30.00|    0.00|  30.00
         |    0.00|  100.00|    0.00|
         |    0.00|   75.00|    0.00|
---------+--------+--------+--------+
3        |       0|       0|       1|      1
         |    0.00|    0.00|   10.00|  10.00
         |    0.00|    0.00|  100.00|
         |    0.00|    0.00|   50.00|
---------+--------+--------+--------+
Total    |       4|       4|       2|     10
            40.00    40.00    20.00   100.00

     Statistics for Table of R1 by R2

          Test of Symmetry
        ------------------------
        Statistic (S)    2.0000
        DF                    3
        Pr > S           0.5724

          Kappa Statistics
Statistic       Value    ASE  95% Confidence Limits
---------------------------------------------------
Simple Kappa   0.6774 0.1941 0.2970          1.0578
Weighted Kappa 0.6154 0.2347 0.1554          1.0754
          Sample Size = 10
```

Figure 4. Output of SAS Procedure with data set of figure 3.

```
   Simple Kappa Coefficient
 ______________________________
 Kappa                  0.6774
 ASE                    0.1941
 95% Lower Conf Limit   0.2970
 95% Upper Conf Limit   1.0578

     Test of H0: Kappa = 0

 ASE under H0           0.2249
 Z                      3.0123
 One-sided Pr > Z       0.0013
 Two-sided Pr > |Z|     0.0026
        Sample Size = 10

   Weighted Kappa Coefficient
________________________________
Weighted Kappa          0.6154
ASE                     0.2347
95% Lower Conf Limit    0.1554
95% Upper Conf Limit    1.0754

Test of H0: Weighted Kappa = 0

ASE under H0            0.2316
Z                       2.6568
One-sided Pr > Z        0.0039
Two-sided Pr > |Z|      0.0079
       Sample Size = 10
```

Figure 5. SAS Output from Proc. 2.

In order to conduct a significance test on the weighted kappa only and not on the simple kappa, one should replace the third line of Proc. 3 (i.e. "TEST KAPPA;") with "TEST WTKAP;".

3.3) Exact P-values

One should note that the P-values and $\boldsymbol{Z}$ scores of simple and weighted kappa statistics are based on Asymptotic Standard Errors (ASE), which will often be invalid when the number of subjects is small. Fortunately, the FREQ procedure provide the option of computing exact P-values with the "EXACT" statement. The EXACT statement must be followed by one of the following three keywords: AGREE (to obtain exact P-values for simple and weighted kappa coefficients), KAPPA (to obtain exact P-values for simple kappa coefficient), WTKAP (to obtain exact P-values for weighted kappa coefficients). When dealing with two raters and two response categories, the TEST statement of the FREQ procedure will also provide an exact P-value for the McNemar's test of marginal homogeneity.

The statements in example 4 would yield an output that is similar to what was discussed earlier. The sections entitled "Simple Kappa Coefficient" and "Weighted Kappa Coefficient" would appear as shown in figure 6. This figure shows that exact P-values may occasionally be quite different from their asymptotic approximation. The computation of exact P-values does not take too long when the subject sample of small or moderate size. Therefore, we recommend to always specify the EXACT statement, unless the number of subjects is large (approximately 200).

Example 4

```
   PROC FREQ DATA = ClassData;
      TABLES R1*R2/AGREE;
      TEST AGREE;
      EXACT AGREE;
      RUN;
```

Even if the number subjects and response categories is large and the user is still concerned about the accuracy of the asymptotic standard errors, the FREQ procedure provides the Monte Carlo option to the EXACT statement. The Monte-Carlo option instructs the SAS system not to use the network algorithm of Mehta and Patel (1983), but rather to generate several random contingency tables with the same marginal totals as the observed table.

The Monte-Carlo method is used if the MC option is specified with the EXACT statement. By default, SAS® will generate 10,000 random contingency tables. One may increase this number by specifying the option "N=" to the EXACT statement. For example N=100,000 will ask the FREQ procedure to generate the FREQ procedure to generate 100,000 random contingency tables.

```
     Simple Kappa Coefficient
  ----------------------------------
  Kappa (K)               0.6774
  ASE                     0.1941
  95% Lower Conf Limit    0.2970
  95% Upper Conf Limit    1.0578

      Test of H0: Kappa = 0

  ASE under H0            0.2316
  Z                       2.6568
  One-sided Pr > Z        0.0039
  Two-sided Pr > |Z|      0.0079

  Exact Test
  One-sided Pr >= K       0.0095
  Two-sided Pr >= |K|     0.0095

    Weighted Kappa Coefficient
 ------------------------------------
 Weighted Kappa            0.6154
 ASE                       0.2347
 95% Lower Conf Limit      0.1554
 95% Upper Conf Limit      1.0754

 Test of H0: Weighted Kappa = 0

 ASE under H0              0.2316
 Z                         2.6568
 One-sided Pr > Z          0.0039
 Two-sided Pr > |Z|        0.0079

 Exact Test
 One-sided Pr >= K         0.0238
 Two-sided Pr >= |K|       0.0286

        Sample Size = 10
```

Figure 6. SAS Output from Example 4.

The statements of example 5 show how the Monte-Carlo option can be used with the EXACT statement. In this example, we have added another option ALPHA=0.10. Its role is to specify the confidence level used to construct the confidence interval around the Monte-Carlo P-value. Figure 7 shows only the "Simple Kappa Coefficient" section of the output generated by the statements of example 5. The results shown in the "Weighted Kappa Coefficient" section can be interpreted the same way.

The statistics under the title "Monte-Carlo Estimates for the Exact Tests" are the novelty in figure 7. The first three statistics show the Monte-Carlo one-sided P-value under the label "Estimate", and the 95% lower and upper confidence limits for the P-value. The closer the two confidence limits, the more precise the Monte-Carlo P-value. If all Monte-Carlo P-values are positive, then the one-sided and two-sided P-values will be identical, as is the case in example 5. The last number in Figure 7 is the initial seed. This number is useful for obtaining the same results each time the program is run. The SEED option should then be used with the EXACT statement for that purpose. A statement such as the following:

```
    EXACT AGREE/MC N=10000 ALPHA=0.05
                       SEED=49545;
```

will always yield the same results of Figure 7.

Example 5

```
   PROC FREQ DATA = ClassData;
      TABLES R1*R2/AGREE;
      TEST AGREE;
      EXACT AGREE/MC N=10000 ALPHA=0.05;
      RUN;
```

```
     Simple Kappa Coefficient
  ----------------------------------
  Kappa (K)               0.6774
  ASE                     0.1941
  95% Lower Conf Limit    0.2970
  95% Upper Conf Limit    1.0578

      Test of H0: Kappa = 0

  ASE under H0            0.2249
  Z                       3.0123
  One-sided Pr > Z        0.0013
  Two-sided Pr > |Z|      0.0026
```

```
Monte Carlo Estimates for the Exact Test
      One-sided Pr >= K
      Estimate               0.0091
      95% Lower Conf Limit   0.0086
      95% Upper Conf Limit   0.0097

      Two-sided Pr >= |K|
      Estimate               0.0091
      95% Lower Conf Limit   0.0086
      95% Upper Conf Limit   0.0097

      Number of Samples      100000
      Initial Seed            49545
```

Figure 7. SAS Output from Example 5.

4. **Limitations of the FREQ Procedure**

The computation of kappa coefficients (weighted and unweighted) with the FREQ procedure is a welcome addition to the SAS system. However, there are still some severe limitations in this implementation that at times could be very annoying. In addition to having limitations, the FREQ procedure may yield misleading (even wrong) results. Here are some of the most serious problems that one should be aware of:

a) Number of raters is limited to 2.

The FREQ procedure can only compute the extent of agreement between 2 raters. For practitioners interested in computing the extent of agreement between multiple raters, a dedicated SAS program is necessary (see section 5 for a complete discussion).

b) Wrong Results

There is a important assumption underlying the computation of the Kappa statistic with the FREQ procedure that users should know. In fact, the AGREE option assumes that each of the two raters classifies at least one subject into each response category considered in the study. This severe restriction leads to the following two problems:

- If the first rater classifies a subject into each of the response categories and that the second rater classify the subjects into all but one category, then SAS will report the following error message: `"WARNING: AGREE statistics are computed only for tables where the number of rows equals the number of columns."` As a result, the AGREE option of the TABLES statement will be ignored and no kappa statistic will be computed.
- Although both raters may use the same number of categories, each of them can still use some categories that the other has not used. In this case, the FREQ procedure will compute the kappa statistics as if both raters have used the same categories. The obtained results will then be erroneous.

Several authors have proposed a variety of ways to solve these two problems. One method often suggested in the literature is to add fictitious observations to the input file in such a way that there is at least one subject in each category for each rater. The fictitious data will be assigned a very small weight so as to reduce their impact on the kappa statistics. A second approach, discussed by Liu and Hays (1999) and implemented in their SAS macro called “%kappa”, consists of creating the correct (balanced) two-way contingency table from the input data set prior to computing the kappa statistics. The Liu-Hayes amounts to filling empty cells with zeros before computing the kappa statistics. The SAS macro that Liu and Hayes have written to implement their method can be downloaded from the following URL: `"http://www.gim.med.ucla.edu/FacultyPages/Hays/UTILS/WKAPPA.txt"`. Their paper describing their approach can also be downloaded from `"http://www2.sas.com/proceedings/sugi24/Stats/p280-24.pdf"`.

We believe that these two approaches are unnecessarily cumbersome. The simplest and clean approach we have found is due to Crewson (2001). One may download the PDF file of his paper at "http://www2.sas.com/proceedings /sugi26/p194-26.pdf".

Crewson(2001) suggests to modify the input data set slightly by assigning original data to one stratum (e.g. STRATUM=1) and by creating in another stratum (e.g. STRATUM=2) with dummy observations to cover all possible rating scenarios. The TABLES statement of the FREQ procedure will then be defined as follows:

```
TABLES STRATUM*R1*R2/AGREE.
```

As a result, the FREQ procedure will compute the kappa statistics (weighted and unweighted) separately for stratum 1 and stratum 2. One will simply use the results from the stratum of interest. The stratum of dummy observations will contain a number of observations (subjects) that equals the number of categories in the study. All raters classifies each dummy subject into the same category, which is different by subject.

5. Computing Kappa for Multiple-Rater Studies

As mentioned earlier, the SAS® system can only compute the extent of agreement between two raters. When the number of raters is three or more, the now widely-accepted extension of Kappa due to Fleiss (1971) cannot be used as it is not implemented in SAS. However, a SAS macro called "magree.sas" and developed by the SAS institute as a service to SAS users, implements the generalized kappa of Fleiss (1971). This SAS macro can be downloaded from the site "http://ewe3.sas.com/techsup/download /stat/magree.html".

The MAGREE.SAS macro is well documented and can compute Kappa as well as the Kendall's coefficient of concordance if the response variable is numeric. This macro also has the special feature of computing kappa statistics conditionally on the response category. That is for each category, the conditional Kappa is computed as a measure of the extent of agreement between raters with respect to that specific category. The methods for computing these conditional kappa coefficients are also discussed by Fleiss (1971).

The version of the MAGREE.SAS macro (version 1.0) that we were able to obtain does not compute the variance of the conditional Kappa properly as specified by Fleiss (1971). In fact the correct variance formula is given by equation [23] of Fleiss' paper. Its implementation in MAGREE.SAS is erroneous although the standard errors obtained are often close to the correct estimations. However, the variance of the overall Kappa was properly implemented according the formula given by Fleiss (1971).

Gwet (2001) has introduced the AC1 statistic as an alternative method for evaluating the extent of agreement between raters. The AC1 statistic is not vulnerable the well-known paradoxes that make Kappa look bad. Moreover, Gwet(2001) also demonstrated that if the variance is only based on the sampling variability of subjects, then it should be seen as being conditional on the sample of raters. Any statistical inference based on a conditional variance is only applicable to the raters who participated to the reliability experiment. In order to infer to the general population of raters and that of subjects, it is necessary to derive a variance based on the sampling distribution of subjects and that of raters as well. Therefore Gwet (2001) has discussed about the conditional and unconditional variances of the AC1 statistic.

Some of the articles posted on the web site www.stataxis.com give an overview of the AC1 statistic and its effectiveness for measuring the extent of agreement between two or multiple raters. We have developed a SAS macro INTER_RATER.MAC that implements the overall and conditional Kappa statistics of Fleiss (1971) as well as the overall and conditional AC1 statistics. For the AC1 statistics, the macro offers the possibility to compute the standard error conditionally upon the rater sample and the unconditional standard error that would allow for inference to the general population of raters. The INTER_RATER.MAC macro can be downloaded from the address http://www.stataxis.com/files/sas /INTER_RATER.TXT.

This file is well documented and explains well how this SAS macro should be used. The macro is evoked by specifying the following command:

```
Inter_Rater(InputData=,DataType=,
            CategoryFile=,OutFile=);
```

where `InputData` is the input file that can take two forms. It may indicate for each subject and each rater, the category into which the rater classified the subject or it may show the distribution of raters by subject and category. Table 3 is an example of input file that provides classification data where 4 raters classified 29 subjects into 5 possible response categories. For such a data set, one should specify `DataType=C`. If the input file contains one subject variable and 5 category variables indicating the number of raters who classified the subject into the specified category, then one should specify `DataType=A`.

The parameter `CategoryFile=` must point to a one-variable file containing the list of all category identifiers. The category file that corresponds to Table 3 can be created as follows:

```
DATA CatFile;
      INPUT Category @@;
      DATALINES;
1 2 3 4 5
;
```

If the `Inter_Rater` macro is run with the statement

```
%Inter_Rater(InputData=ClassData,
      DataType=c,VarianceType=u,
      CategoryFile=CatFile,OutFile=a2);
```

then the output will contain the results described in Tables 4, 5, and 6. Table 7 describes the content of the output file `a2`. This SAS file can be used for further processing.

All the variables in the data sets can be of any type (character or numeric). This SAS macro will also handle the case where not all response categories are used by each rater.

Table 3. Classification of 29 subjects into 5 categories.

Subject	R1	R2	R3	R4
1	5	5	5	5
2	3	1	3	1
3	5	5	5	5
4	3	1	3	1
5	5	5	4	5
6	1	2	3	3
7	3	1	1	1
8	1	1	1	3
9	3	3	4	4
10	1	3	1	1
11	5	5	5	5
12	1	1	1	1
13	1	1	1	1
14	1	1	1	1
15	3	3	4	3
16	1	3	3	4
17	4	4	5	5
18	5	5	5	5
19	5	3	3	3
20	3	2	3	3
21	5	3	5	5
22	3	3	3	4
23	1	1	1	1
24	1	1	1	1
25	1	1	3	3
26	3	3	3	1
27	3	3	1	1
28	3	3	1	1
29	3	3	2	5

```
                    Table 4.
           INTER_RATER macro (v 1.0)
Kappa statistics: conditional and unconditional analyses
                    Standard
Category    Kappa     Error       Z      Prob>Z

1          0.52724    0.21791   2.41957  0.00777
2         -0.02655    0.23766  -0.11171  0.54447
3          0.16661    0.20781   0.80175  0.21135
4          0.10494    0.18859   0.55645  0.28895
5          0.73561    0.18982   3.87531  0.00005
Overall    0.41035    0.05203   7.88715  0.00000
```

```
                    Table 5.
            INTER_RATER macro (v 1.0)
AC1 statistics: conditional and unconditional analyses
    Inference based on conditional variances of AC1
             AC1     Standard
Category  Statistic    Error       Z      Prob>Z

1           0.63316   0.23680   2.67379  0.00375
2          -0.21636   0.00000         .        .
3           0.30963   0.13032   2.37596  0.00875
4          -0.01363   0.14077  -0.09686  0.53858
5           0.75049   0.36171   2.07484  0.01900
Overall     0.48969   0.06870   7.12822  0.00000
```

```
                    Table 6.
            INTER_RATER macro (v 1.0)
AC1 statistics: conditional and unconditional analyses
   Inference based on unconditional variances of AC1
             AC1     Standard
Category  Statistic    Error       Z      Prob>Z

1           0.63316   0.28358   2.23275  0.01278
2          -0.21636   0.00000         .        .
3           0.30963   0.16509   1.87559  0.03036
4          -0.01363   0.14563  -0.09363  0.53730
5           0.75049   0.38906   1.92897  0.02687
Overall     0.48969   0.21936   2.23235  0.01280
```

```
                                 Table 7.
          -----Alphabetic List of Variables and Attributes-----
 #  Variable        Type  Len  Pos  Label
14  AC1Statistic    Num     8  104  AC1 estimate
11  AC1_CondVar     Num     8   80  Conditional variance of AC1
21  AC1_UcondVar    Num     8  160  Unconditional variance of AC1
20  Cr              Num     8  152  Cr factor for unconditional variance
 9  FleissVar       Num     8   64  Kappa variance according to Fleiss
13  KappaStat       Num     8   96  Kappa estimate
19  P2aq            Num     8  144  P2a probability
 7  PaQ             Num     8   48  Agreement probability conditional on q
 4  Pe_gamma        Num     8   24  AC1 Chance-agreement probability
 5  Pe_kappa        Num     8   32  Kappa Chance-agreement probability
 8  PiHatQ          Num     8   56  Classification probability in category q
18  PvalAC1         Num     8  136  AC1 conditional P-value
24  PvalAC1_U       Num     8  184  AC1 unconditional P-value
17  PvalKappa       Num     8  128  Kappa conditional P-value
 2  Raters          Num     8    8  Number of raters
22  StdErrAC1_U     Num     8  168  AC1 unconditional standard error
12  StdErrAC1cond   Num     8   88  AC1 conditional standard error
10  StdErrKappa     Num     8   72  Kappa standard error
 1  Subjects        Num     8    0  Number of subjects
16  Z_AC1           Num     8  120  AC1 conditional Z score
23  Z_AC1_U         Num     8  176  AC1 unconditional Z score
15  Z_Kappa         Num     8  112  Kappa Z score
 3  q               Num     8   16  Category number
 6  rBarQ           Num     8   40  Average number of raters in category q
```

6. Concluding Remarks

The AGREE option provided in the FREQ procedure is a welcome addition to the SAS system. It allows SAS users to easily obtain simple and weighted Kappa statistics using an already widely-used procedure. Standard errors and P-values can also be obtained for statistical inference. The FREQ procedure also provides the option to compute exact P-values using either the network algorithm or using the Monte-Carlo simulation approach.

The implementation of Kappa in SAS is unfortunately limited to two raters and occasionally produces errors when some categories are not used by one rater. The `MAGREE.SAS` SAS macro developed at the SAS Institute implements the generalized Kappa statistics proposed by Fleiss (1971). In addition to producing the overall kappa, this macro can compute kappa coefficients conditionally on a specific response category. These conditional statistics allow researchers to evaluate the propensity of raters to agree on a specific category. However, the first version of the `MAGREE.SAS` macro does not implement the standard error of the conditional kappa properly.

Gwet (2001) strongly recommended the use of the AC1 statistic in order to evaluate the extent of agreement between raters. The computation of this statistic and the associated standard error can be carried out using the `INTER_RATER.MAC` SAS macro that can be downloaded from the internet. It computes the overall kappa and AC1 statistics as well as their conditional versions with respect to specific categories. Unconditional standard errors allowing inference to the universe of raters can also be calculated for the AC1 statistic. This macro could have been simplified substantially using the IML procedure of SAS. Because this procedure must be licensed separately, it is not always present in all SAS systems. The `INTER_RATER.MAC` macro can therefore run in any SAS system running BASE SAS.

7. References

Bennet et al. (1954). Communications through limited response questioning. *Public Opinion Quarterly*, **18**, 303-308.

Cicchetti, D.V. and Feinstein, A.R. (1990). High agreement but low Kappa: II. Resolving the paradoxes. *Journal of Clinical Epidemiology*, **43**, 551-558.

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

Crewson, P.E. (2001). A Correction for Unbalanced Kappa Tables. `http://www2.sas.com/proceedings/sugi26/p194-26.pdf`

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters., *Psychological Bulletin*, **76**, 378-382.

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

Gwet, K. (2002). *Handbook of Inter-Rater Reliability.* STATAXIS Publishing Company.

Liu, H. and Hays, R.D. (1999). Measurement of Interrater Agreement: A SAS/IML Macro Kappa Procedure for Handling Incomplete Data. *Proceedings of the Twenty-Fourth Annual SAS*

Users Group International Conference, April 11-14, 1999, 1620-1625.

Mehta, C.R. and Patel, N.R. (1983). A Network Algorithm for Performing Fisher's Exact Test in $\boldsymbol{r} \times \boldsymbol{c}$ Contingency Tables. *Journal of the American Statistical Association*, **78**, 427-434.

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

Zwick, R. (1988). Another look at Interrater Agreement. *Psychological Bulletin*, **103**, 374-378.

CONTACT

Kilem Gwet, Ph.D.
Sr. Statistical Consultant,
STATAXIS Consulting
15914B Shady Grove Road, #145,
Gaithersburg, MD 20877, U.S.A.

E-mail: kilem62@yahoo.com
Fax: 301-947-2959.