Paper 186-2007

Disagreement on Agreement: Two Alternative Agreement Coefficients

Emily Blood and Kevin F. Spratt, Dartmouth Medical School, Hanover, NH

ABSTRACT

Everyone agrees there are problems with currently available agreement coefficients. Cohen's weighted Kappa does not extend to multiple raters, and does not adjust for both chance agreement and misclassification errors. There is not, however, agreement about what to do about these issues. In an effort to solve some of the problems with current coefficients, Kilem Gwet theoretically derived two alternatives to current agreement coefficients. The first- and second-order agreement coefficients, AC1 and AC2, are presented in his *Handbook of Inter-Rater Reliability: How to Measure the Level of Agreement Between Two or Multiple Raters* (2001). These coefficients adjust for chance agreement and both chance agreement and misclassification, respectively. These are useful tools for reliability studies, however, computation of the statistics and their conditional and unconditional variances beyond the two-rater, two-category case is nontrivial. The formulation of Gwet's statistics is summarized and a SAS® macro is introduced that allows SAS® users to easily compute the statistics and their variances. An example of the use of the macro is also provided.

INTRODUCTION

The increased emphasis of accountability in education, business, medicine and government over the last decade has raised the publics' consciousness about the importance of knowing the reliability and validity of the information used to establish this accountability. If the numbers form the basis for judging accountability, how much trust can we place in the numbers? Unfortunately, there is no unified theory for reliability. For continuous outcomes generalizability provides the most robust framework for evaluating reliability. For dichotomous outcomes, a number of procedures are available with Cohen's Kappa (Cohen 1960,1968) statistic prominent among them. However, statisticians, social scientist and increasingly biostatisticians have become increasing aware and wary of Kappa as a measure of reliability and agreement for a number of reasons. As summarized by Uebersax (2002).

1. "Kappa is not really a chance-corrected measure of agreement
2. Kappa is an omnibus index of agreement. It does not make distinctions among various types and sources of disagreement.
3. Kappa is influenced by trait prevalence (distribution) and base-rates. As a result, Kappas are seldom comparable across studies, procedures, or populations (Feinstein & Cicchetti 1990, Thompson & Walter 1988).
4. Kappa may be low even though there are high levels of agreement and even though individual ratings are accurate. Whether a given Kappa value implies a good or a bad rating system or diagnostic method depends on what model one assumes about the decision making of raters (Uebersax 198[7]).
5. With ordered category data, one must select weights arbitrarily to calculate weighted Kappa (Maclure & Willett 1987).
6. Kappa requires that two rater/procedures use the same rating categories. There are situations where one is interested in measuring the consistency of ratings for raters that use different categories (e.g., one uses a scale of 1 to 3, another uses a scale of 1 to 5). "

Because of these known and documented problems, guidelines such as those by Landis and Koch (1977) that purport to categorize ranges of Kappa as "good," "fair" and "poor" are inappropriate and should not be used.

Recently Gwet (2001) has proposed another approach for evaluating agreement that addresses and may resolve many of the problems affecting Kappa. First, coefficients can be calculated across more than two raters (although there is a generalized Kappa that does allow this). Second, the agreement coefficient can be adjusted for both chance agreement and misclassification errors, which Kappa statistics do not allow.

INTRODUCTION TO AC1 AND AC2

Gwet has proposed two new agreement coefficients (Gwet 2001). The first is for use with any number of raters using a categorical rating system to rate objects. The second is for use with any number of raters using an ordered categorical rating system to rate objects. The first agreement coefficient is called the first-order agreement coefficient or AC1 statistic. This agreement coefficient adjusts the overall agreement probability for chance agreement. Chance agreement occurs when raters agree on a rating due to one or both of the raters giving a random rating. A random rating occurs when a rater is not certain how to classify an object. This can occur when the object's characteristics do not match the rating instructions. Chance agreement can inflate the overall agreement probability and should not contribute to a measure of actual agreement between raters. Therefore, as is done with the Kappa statistic, Gwet has adjusted for chance-agreement in his AC1 statistic. The AC1 between two or multiple raters "is defined as the conditional probability that two randomly selected raters agree given that there is no agreement by chance." (Gwet 2001)

The second-order agreement coefficient, or AC2 statistic, also adjusts for chance-agreement. In addition, it adjusts for misclassification errors. With an ordered categorical scale, some disagreements between raters are considered more serious than others. If two raters have similar assessments of an object, but classify the objects differently, this is considered a classification error. If two categories of a rating scale are similar, the probability that an independent reviewer would reclassify the patient originally classified into one category into the other category based on the original reviewers assessment would be high. If the categories are very different, the probability for such a reclassification would be low. Misclassification errors can bias the inter-rater reliability estimate. If the probabilities of misclassification are provided, the second-order agreement coefficient can be adjusted for chance-agreement as well as misclassification errors.

We have written a SAS macro, AC1AC2.mac[1] that computes the AC1 and AC2 statistic and their conditional and unconditional variances.

DETAILS ON AC1 AND AC2

GENERAL FORMULATION OF A CHANCE-ADJUSTED AGREEMENT COEFFICIENT

The general method for adjusting for chance-agreement in an agreement coefficient is to get an estimate of the overall probability of agreement and condition that on there being no chance agreement. The overall probability of agreement is expressed as P_a. The chance agreement is expressed as P_e. As Gwet describes, the conditional probability of two raters agreeing conditional on there being no chance agreement is:

P(raters agree| NOT(raters agree & rating was by chance rating)).

Which is equal to the following by the definition of conditional probability:

P(raters agree & NOT(raters agree & rating was by chance))/
P(NOT(raters agree & rating was by chance)).

Since the event 'raters agree' can be partitioned into events 'raters agree & rating was by chance' and 'raters agree & rating wasn't by chance', we can write the following:

P(raters agree)-P(raters agree & rating was by chance)/
1-P(raters agree & rating was by chance).

Now, let P_e be the probability that raters agree and the rating is by chance and P_a the probability that agreement is by chance or not by chance. Therefore, the general formula for chance-adjusted agreement is:

$$AC = \frac{P_a - P_e}{1 - P_e}$$

The way P_a and P_e will be calculated depends whether AC1 or AC2 is calculated, both will be derived later.

CHANCE AGREEMENT

Chance agreement occurs when one or both raters rate an object based on a random rating. Cohen's Kappa statistic and Gwet's agreement coefficients both adjust for this type of agreement to avoid inflating the agreement probability with agreements that do not reflect true intentional agreement between the raters. Since it cannot be known which agreements between raters are real and which are chance-agreements, the probability of chance agreement must be estimated (Gwet 2001).

MULTIPLE RATERS AND MULTIPLE-LEVEL SCORING SCALES

The formula for the AC1 and AC2 are given below for the most general situation, that of more than two raters and a rating scale with greater than two categories.

FIRST-ORDER AGREEMENT COEFFICIENT

We need to estimate both overall agreement probability, P_a, and the chance-agreement probability, P_e, for AC1. To do this we introduce some notation, which generally matches that provided by Gwet. We let r_{iq} indicate the number of raters who classified the ith object into the qth category. The index i ranges from 1 to n and q ranges from 1 to Q, where n is the number of objects rated and Q is the number of categories in the rating scale. We let r indicate the total number of raters in our study. For use in our $p_{e\gamma}$ formula, we need to calculate π_q, the probability that a rater classifies an object into category q. The γ in the $p_{e\gamma}$ formula indicates that this calculation takes into account the probability of a random rating (γ). These formulas to the right allow calculation of AC1.

$$\pi_q = \frac{1}{n}\sum_{i=1}^{n}\frac{r_{iq}}{r}$$

$$p_{e\gamma} = \frac{1}{Q-1}\sum_{q=1}^{Q}\pi_q(1-\pi_q)$$

$$p_a = \frac{1}{n}\sum_{i=1}^{n}\left\{\sum_{q=1}^{Q}\frac{r_{iq}(r_{iq}-1)}{r(r-1)}\right\}$$

$$AC1 = \frac{p_a - p_{e\gamma}}{1-p_{e\gamma}}$$

MISSCLASSIFICATION PROBABILITIES

The seriousness of disagreement is represented by the probability that an independent rater seeing the original rater's assessment of the object could then classify the object

[1] http://mcrc.hitchcock.org/SASMacros/Agreement/AC1AC2.txt

differently. If there is a re-classification from one category to another, the original rater is said to have made a classification error. These types of errors can then be adjusted for in the computation of the agreement coefficient statistic. This is a method for weighting disagreements differently. This most often is applicable when the rating scale is ordinal and a disagreement by one or two categories would be less serious than a disagreement by larger amounts. Disagreement severity is captured in the misclassification probabilities, or *β*'s. These *β*'s are conditional probabilities that describe, given an object was originally classified into one category, the probability that the same object would be reclassified into another category. For example, on a 5-point rating scale, given an object is originally classified into category 1, the probabilities that this object would be reclassified into categories, 1,2,3,4 and 5 are given by $\beta_{1|1}$, $\beta_{2|1}$, $\beta_{3|1}$, $\beta_{4|1}$, and $\beta_{5|1}$, respectively. These five probabilities must sum to 1 since reclassification must be to one of the five possible categories.

Additionally, five probabilities must be determined for objects originally classified into category 2, 3, 4, and 5. In Gwet's notation, these probabilities are represented in a matrix with the initial rating determined by the column and the new rating determined by the row. Therefore, the columns sum to 1 and the matrix looks like the one to the right for a five-category scale.

$$\begin{bmatrix} \beta_{1|1} & \beta_{1|2} & \beta_{1|3} & \beta_{1|4} & \beta_{1|5} \\ \beta_{2|1} & \beta_{2|2} & \beta_{2|3} & \beta_{2|4} & \beta_{2|5} \\ \beta_{3|1} & \beta_{3|2} & \beta_{3|3} & \beta_{3|4} & \beta_{3|5} \\ \beta_{4|1} & \beta_{4|2} & \beta_{4|3} & \beta_{4|4} & \beta_{4|5} \\ \beta_{5|1} & \beta_{5|2} & \beta_{5|3} & \beta_{5|4} & \beta_{5|5} \end{bmatrix}.$$

For any rating scale, the diagonal elements will likely by large. With no probability of misclassification, the matrix would be the identity matrix. In fact, with the identity matrix as the misclassification matrix, the AC1 and AC2 statistics are identical. This matrix of β's must be supplied to the macro by the user and should be based on prior knowledge of the rating scale.

The example beta matrix at the right might be applicable when the likelihood of misclassification becomes less as the distance between categories becomes greater. Note that: 1) in each column the probabilities sum to 1.0; 2) in general the probably of correct original classification (probabilities on the diagonal) are .8, but is .7 for the middle category, where options of misclassification to lower or higher levels are greater; and 3) probabilities of misclassification diminish, but do not disappear, as the distance of the category moves farther away from the diagonal element.

.80	.06	.05	.02	.01
.10	**.80**	.1	.05	.03
.06	.07	**.70**	.07	.06
.03	.05	.1	**.80**	.1
.01	.02	.05	.06	**.80**

From the beta matrix of values, additional probabilities must be computed. These are the alpha probabilities. Gwet defines "for two arbitrary categories *q* and *k*, the probability $\alpha_{a|qk}$ for a reviewer to reclassify into the same category an object originally classified into categories *q* and *k* by 2 raters". That is, the sum across all *Q* categories, of the probability of reclassifying each possibly combination of ratings into a given category. Assuming classification errors are independent he then states the "fundamental equation of classification errors" as $\alpha_{a|qk} = \sum_{l=1}^{Q} \beta_{l|q}\beta_{l|k}$. The *α*'s are used in the computation of the overall agreement probability, p_a, when misclassification exists since they represent the conditional probability of agreement for any pair of rater classifications of an object. For example, $\alpha_{a|12}$ (the alpha of agreement given classifications of 1 and 2) represents the probability that a pair of raters that classify an object into categories 1 and 2, respectively, actually agree on the classification of this object. We would expect the $\alpha_{a|qk}$ to be larger when categories *q* and *k* are closer. The alpha probabilities and the likelihood of each possible combination of rater classifications for a pair of raters are combined to compute the overall agreement probability for a pair of raters. The alpha probability matrix derived from the beta matrix is computed in the macro.

SECOND-ORDER AGREEMENT COEFFICIENT

Again, the overall agreement probability, P_a, and the chance-agreement probability, P_e, must be estimated for AC2. Since AC2 adjusts for both chance-agreement as well as misclassification errors, the calculation of the quantities is done differently than for AC1. In the formulas below, *l* is an index for scale category and ranges from 1 to *Q*. The meaning of the other letters, unless otherwise specified, is the same as that above in AC1 formulas. AC2 is defined as the "bias-adjusted" conditional probability that two randomly chosen raters agree given that there is no agreement by chance". The formulas necessary to calculate the AC2 statistic are shown below.

$$\pi_l = \frac{1}{n}\sum_{i=1}^{n}\frac{r_{il}}{r} \qquad \alpha_{a|ql} = \sum_{k=1}^{Q}\beta_{k|q}\beta_{k|l} \qquad p_{e\gamma}^{'} = \frac{1}{Q-1}\sum_{q=1}^{Q}\pi_q^{'}(1-\pi_q^{'})$$

$$\pi_q^{'} = \sum_{l=1}^{Q}\beta_{q|l}\pi_l \qquad p_a^{'} = \frac{1}{n}\sum_{i=1}^{n}\left\{\sum_{q=1}^{Q}\alpha_{a|qq}\frac{r_{iq}(r_{iq}-1)}{r(r-1)} + \sum_{q\neq l}^{Q}\sum_{q\neq l}^{Q}\alpha_{a|ql}\frac{r_{iq}r_{il}}{r(r-1)}\right\} \qquad AC2 = \frac{p_a^{'} - p_{e\gamma}^{'}}{1-p_{e\gamma}^{'}}$$

CONDITIONAL VARIANCE AND UNCONDITIONAL VARIANCE

The difference between conditional and unconditional variances, as detailed by Gwet, reflect differences in agreement statistic interpretation. The conditional variance is appropriate when the inference is only on the given sample of raters. In this case, the inferences based on the conditional variability of the AC statistic cannot be generalized to the universe of raters. If a researcher is only interested in the set of raters currently under investigation, then the conditional variance can be used to make inferences about this set. This could occur, for example, if the researcher is interested in the agreement among raters at a single school, or hospital or agency. In this case, a study of agreement among all raters at this institution would be sufficient to generalize to this population of interest and the conditional variance could be appropriately used. However, it is quite common for decision makers to want to generalize the inferences regarding agreement levels observed within a particular setting to similar situations that have yet to be evaluated. The unconditional variance, which takes into account the variability associated with selecting a set of raters from the universe of raters, is appropriate when inferences are to be generalized beyond the given sample. If a researcher is interested in agreement for a test used at several institutions and a study on agreement selects a random sample of raters from among all raters at these institutions, the *un*conditional variance would be appropriate to make inferences on the whole population of raters at the institutions. *Thus, if you want to understand the agreement between the three raters in your study, conditional variance estimates should be used to estimate a confidence interval around the obtained agreement statistic. If, on the other hand, you would like to generalize these results to any of the clinicians who might be asked to classify that patient, the unconditional variance estimate should be used.*

CONDITIONAL VARIANCE OF FIRST-ORDER AGREEMENT COEFFICIENT

In order to give the formula for the conditional variance of the AC1 statistic, additional notation is needed. An overall probability of agreement, p_a, is determined for each object in the study. This quantity is called p_{ai} and is calculated with the formula to the right.

$$p_{ai} = \sum_{i=1}^{Q} \frac{r_{iq}(r_{iq}-1)}{r(r-1)}$$

Assuming we are using all raters in the population of interest, the AC1 statistic calculated with p_{ai} rather than p_a, is denoted as $\mathrm{K}_{\gamma i}$, while the AC1 statistic calculated with p_a (as in the previous section) is denoted as K_{γ}. Additionally, the sampling fraction is denoted *f*. The sampling fraction is the proportion of objects in the current sample to objects in the universe of objects. In the macro, the sampling fraction has always been assumed to be negligible. The estimate of the conditional variance of the AC1 statistic is calculated using the formulas to the right.

$$S_{\gamma}^2 = \frac{1}{n-1}\sum_{n=1}^{n}\left(\mathrm{K}_{\gamma i} - \mathrm{K}_{\gamma}\right)^2$$

$$CV(AC1) = \frac{1-f}{n} S_{\gamma}^2 .$$

UNCONDITIONAL VARIANCE OF FIRST-ORDER AGREEMENT COEFFICIENT

In order to compute the estimate of the unconditional variance of the AC1 statistic, one more quantity needs to be introduced. This is the number of sample raters who have classified object *i* into category *q* and object *j* into category *l*, denoted $m_{iq\bullet jl}$. In the formulas used here, *r* is assumed to be the number of raters in the sample. Assuming we are sampling from the population of raters, the AC1 statistic calculated with p_{ai} rather than p_a, is denoted as $\mathrm{K}_{\gamma i}$, while the AC1 statistic calculated with p_a (as in the previous section) is denoted as K_{γ}. With the formulas to the right the unconditional variance of the AC1 statistic can be estimated.

$$p_{2a} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{q=1}^{Q}\sum_{l}^{Q}\frac{m_{iq\bullet jl}(m_{iq\bullet jl}-1)}{r(r-1)}$$

$$cr = [r(r-1)(1-p_{e\gamma})^2]^{-1}$$

$$UV(AC1) = \frac{1-f}{n}S_{\gamma}^2 + cr\left\{p_{2a} + \frac{1-f}{n}\left(p_a - p_{2a}\right)\right\}.$$

CONDITIONAL VARIANCE OF SECOND-ORDER AGREEMENT COEFFICIENT

The conditional variance of AC2 is estimated in the same way that the conditional variance for AC1 is estimated, however, $\mathrm{K}'_{\gamma i}$ and K'_{γ} are in the formulas in place of $\mathrm{K}_{\gamma i}$ and K_{γ}. Where $\mathrm{K}'_{\gamma i}$ is the second-order agreement coefficient computed for a single object, computed: $\mathrm{K}'_{\gamma i} = \dfrac{p'_{ai} - p'_e}{1 - p'_e}$, with

$p'_{ai} = \sum_{q=1}^{Q} \alpha_{a|qq} \frac{r_{iq}(r_{iq}-1)}{r(r-1)} + \sum_{q\neq l}^{Q}\sum_{q\neq l}^{Q} \alpha_{a|ql} \frac{r_{iq}r_{il}}{r(r-1)}$. The conditional variance of the AC2 statistic, conditional on the set of raters in the sample, is then estimated as follows:

$$S'^{2}_{\gamma} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\mathrm{K}'_{\gamma i} - \mathrm{K}'_{\gamma}\right)^2$$

$$CV(AC2) = \frac{1-f}{n} S'^{2}_{\gamma}$$

UNCONDITIONAL VARIANCE OF SECOND-ORDER AGREEMENT COEFFICIENT

To give the formula for the estimator of the unconditional variance of the AC2 statistic, some additional formulas are needed. The formulas are:

$$p'_{2a\bullet ij} = \sum_{q=1}^{Q}\sum_{l=1}^{Q} \alpha_{a|qq}\alpha_{a|ll} \frac{m_{iq\bullet jl}(m_{iq\bullet jl}-1)}{r(r-1)} + \sum_{q\neq l}^{Q}\sum_{q\neq l}^{Q}\sum_{k\neq h}^{Q}\sum_{k\neq h}^{Q} \alpha_{a|ql}\alpha_{a|kh} \frac{m_{iq\bullet jk}m_{il\bullet jh}}{r(r-1)}$$

$$p'_{2a} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} p'_{2a\bullet ij}$$

$$p''_{a} = \frac{1}{n}\sum_{i=1}^{n}\left\{\sum_{q=1}^{Q}\alpha^2_{a|qq}\frac{r_{iq}(r_{iq}-1)}{r(r-1)} + \sum_{q\neq l}^{Q}\sum_{q\neq l}^{Q}\alpha^2_{a|ql}\frac{r_{iq}r_{il}}{r(r-1)}\right\}$$

Then, the unconditional variance of the AC2 statistic is estimated as follows:

$$cr' = [r(r-1)(1-p'_{e\gamma})^2]^{-1}$$

$$UV(AC2) = \frac{1-f}{n}S'^{2}_{\gamma} + cr'\left\{p'_{2a} + \frac{1-f}{n}\left(p''_{a} - p'_{2a}\right)\right\}$$

Gwet's notation is changed slightly between the conditional and unconditional variances to make clear that the first is being conditioned on the assumption that the current set of raters is the only one to which inference is to be made and the second is not conditioned on this assumption and inference can be made to all possible raters (e.g. $\mathrm{K}'_{\gamma i|r}$ and $\mathrm{K}'_{\gamma|r}$ versus $\mathrm{K}'_{\gamma i}$ and K'_{γ}). We have not done the same here since the notation does not change how the calculations are performed. Despite this difference in notation, however, we make the same assumptions as Gwet regarding conditioning on the set of raters here.

Gwet notes that the variances given here are only correct when the data are balanced (i.e. All raters have rated all objects). Therefore the macro does not run on unbalanced data. Objects not evaluated by all raters cannot be considered when determining AC1 or AC2 statistics and should be removed from the input data set. Confidence intervals assuming a normal distribution and using the calculated unconditional variances were quite similar to bootstrap percentile-based confidence intervals generated with 1000 iterations.

DETAILS ON SAS MACRO

MACRO REQUIREMENTS

With a general explanation of the concepts and formulas necessary to generate the AC1 and AC2 statistic complete, details are now given for the use of the SAS macro, AC1AC2.mac. In order to output the AC1 and AC2 statistics along with estimates of their conditional and unconditional variances, the macro needs: 1) a classification dataset; 2) the number of categories in the rating scale, and 3) a misclassification probability matrix (beta matrix). The classification dataset needs to be of the form where each row is one object being rated and each column is the score given by one rater. The observations in the dataset need to be positive integers (i.e. greater than or equal to 1). The macro treats the first numeric variable of the dataset as the ratings from the first reader. Therefore, if there is a variable that denotes a subject ID, this variable must be character. The number of categories in the rating scale is just entered as an integer. The beta matrix must be a dataset, with square dimensions determined by the number of categories of the rating scale. The columns determine the initial category and the rows determine the revised category. As indicated previously when summarizing the properties of the beta matrix, the columns of this dataset must sum to 1. The macro generates errors if any of the three required parameters are missing, if the dimensions of the beta matrix are not correct or if the sums of the columns in the beta matrix are not all 1 (within 0.0000001), or if the classification matrix contains missing values.

EXAMPLE FROM GWET'S BOOK

Gwet gives an example dataset on page 129 in the form necessary for this macro. The dataset is part of exercise 1 at the end of Chapter 5. The dataset represents the ratings of six psychologists on the diagnosis of mental illness. For this dataset, and beta matrix, the AC1, AC2, and conditional and unconditional variances are estimated. The beta matrix is one given in Gwet's exercise. The SAS code containing the datasets and the macro call is as follows:

```
* CLASSIFICATION DATA *;
DATA one;
 INPUT r1 r2 r3 r4 r5 r6;
CARDS;

 4 4 4 4 4 4
 2 2 5 2 5 5
 3 3 5 2 3 3
 5 5 5 5 5 5
 2 4 2 4 4 2
 1 3 3 3 1 3
 3 5 3 3 5 3
 1 1 3 3 3 4
 4 4 4 4 1 1

 5 5 5 5 5 5
 1 4 4 4 4 4

 1 4 2 4 4 4
 2 3 2 2 3 3
 4 1 4 4 4 4
 2 2 4 4 4 5
 3 3 5 3 3 3
 5 5 1 1 1 4
 1 1 1 1 2 1

 2 2 4 4 4 4
 1 3 3 5 5 5
 5 5 5 5 5 5
 4 4 2 4 4 4
 5 2 5 5 4 2
 1 4 4 4 1 4
 5 4 4 4 4 1
 2 4 2 2 2 2
 1 5 1 1 1 5

 4 2 4 4 4 2
 1 3 3 3 3 3

 5 5 5 5 5 5
;
RUN;
* BETA MATRIX *;
DATA beta;

INPUT cat1 cat2 cat3 cat4
cat5 ;
CARDS;
.9  .9 .2 .1 0
.05 .1 .8 .7 0
.03  0  0 .1 0
.01  0  0 .1 0
.01  0  0  0 1
;
RUN;
%include "[macro
pathname]";
%AC1AC2(dataset=ONE,
Numcategories=5,
betadata=BETA)
```

For this data, the AC1 result is 0.45 with conditional variance 0.0030 and unconditional variance 0.020. The AC2 result is 0.36 with conditional variance 0.0028 and unconditional variance 0.012. For AC1 and AC2, the p_a and p_e are given as well. For AC1 p_a is 0.56 and p_e is 0.20. For AC2 p_a is 0.47 and p_e is 0.17.

CONCLUSIONS

Gwet has proposed two agreement coefficients that can be used in place of or along with Cohen's Kappa statistic. There is a need for a coefficient that extends to multiple raters and can incorporate weighting disagreements based on the seriousness of the disagreement. Gwet's coefficients can be used for both of these purposes. The following SAS macro can be used to easily calculate the statistics proposed by Gwet along with their conditional and unconditional variances.

Although the AC1 and AC2 statistics are about five years old now, they remain infants in the statistical world, especially since so few people have been exposed to them. With greater usage will come greater scrutiny, and with greater scrutiny may come identification of problems inherent in these statistics. Therefore, as is always the case with new statistics, caution should be exercised in their use and further examination should occur before they are adopted as the standard.

This macro is available at: http://mcrc.hitchcock.org/SASMacros/Agreement/AC1AC2.TXT

REFERENCES

1. Cohen, J., *A coefficient of agreement for nominal scales.* Educational and Psychological Measurement, 1960. **20**: p. 37-46.
2. Cohen, J., *Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit.* Psychological Bulletin, 1968. **70**(4): p. 213-220.
3. Uebersax J. 2002. "Kappa Coefficients: A Critical Appraisal. http://ourworld.compuserve.com/homepages/jsuebersax/kappa.htm. (March 1, 2007)
4. Feinstein AR. Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes [see comments]. Journal of Clinical Epidemiology. 43(6):543-9, 1990.
5. Thompson WD. Walter SD. A reappraisal of the kappa coefficient. Journal of Clinical Epidemiology. 41(10):949-58, 1988.
6. Thompson WD. Walter SD. Kappa and the concept of independent errors. Journal of Clinical Epidemiology, 1988, 41, 969-70.
7. Uebersax JS. Measuring diagnostic reliability: Reply to Spitznagel and Helzer (letter). Archives of General Psychiatry, 1987, 44, 193-194.
8. Uebersax, J. S. (1987). Diversity of decision-making models and the measurement of interrater agreement. Psychological Bulletin, 101, 140-146. Landis, J.R. and G.G. Koch, *The measurement of observer agreement for categorical data.* Biometrics, 1977. **33**(1): p. 159-74.

9. Maclure M, Willett WC. Misinterpretation and misuse of the kappa statistic. American Journal of Epidemiology. 126(2)161-9, 1987 Aug. [dissenting letter and reply appears in Am J Epidemiol 1888 Nov.;128(5)1179-81].
10. Landis, .JR. and G.G. Koch, *The measurement of observer agreement for categorical data.* Biometrics, 1977. **33**(1): p. 159-74.
11. Gwet, K. Handbook of Inter-Rater Reliability: How to measure the level of agreement between two or multiple raters. 2001: Stataxis Publishing Company, PO box 120185 Gaithersburg, MD. 309.

CONTACT INFORMATION

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

Emily A. Blood or Kevin F. Spratt
Department of Orthopedic Surgery
Dartmouth Medical School
One Medical Center Drive
Hanover, NH 03756
Work Phone: 603-653-6019
E-mail: Emily.a.blood@dartmouth.edu or Kevin.F.Spratt@dartmouth.edu

APPENDIX

```
/*************************************************************************************
AC1AC2 MACRO DOCUMENTATION
Version 1.0
Authors:  Emily A. Blood, M.S.    Dept of Orthopaedic Surgery, Dartmouth Medical School
          Kevin F. Spratt, PH.D.  Dept of Orthopaedic Surgery, Dartmouth Medical School

Users with comments, suggestions, or who have identified problems should
contact Emily Blood via e-mail at Emily.A.Blood@Dartmouth.edu

DISCLAIMER:
-------------------------------------------------------------------------------------
!!!No guarantee as to suitability or accuracy is given or implied. User uses this !!!
!!! code entirely at his/her own risk!!!
-------------------------------------------------------------------------------------
OVERVIEW
Reference:
This macro is based on the work presented by Kilem Gwet in
Gwet, Kilem. 2001.  Handbook of Inter-Rater Reliability. Stataxis Publishing Company,
Gaithersburg, MD.

Macro Function:
The AC1AC2 macro generates two agreement coefficients (AC1 and AC2). As defined by Gwet, AC1 is
overall agreement probability conditioned on the absence of chance agreement and AC2 is the
overall agreement probability adjusted for classification errors and conditioned on the absence
of chance agreement. The adjustment for classification errors is based on probabilities
associated with the likelihood that a category chosen on one occasion by a generalized rater is
likely to change on a subsequent occasion.

MACRO INPUT:
The macro has three parameters:

  %AC1AC2(dataset=,numcategories=,betadata=)

**dataset is the SAS data set with each object of classification making up a separate row and
each rater's classification for each object is a separate variable. Thus if there are 3 raters
and 4 objects the required data structure is:

          r1 r2 r3    (variable names, not in the cards statement)

           1  2  2
           2  2  1
           2  1  2
           2  2  2

Note that an object ID for each row is not included in the data set. The variable names for the
raters must conform to SAS variable name conventions but otherwise can be anything at all. The
```

variable names for the raters must conform to SAS variable name conventions but otherwise can be anything at all. The ratings, however, must be integers greater than 0.

**numcategories requires that you enter an integer reflecting the number of possible categories that a rater has to choose from (NOT the number of categories used). In the example above numcategories would equal 2.

**Betadata is a data set that reflects the transition probabilities associated with a generalized rater changing their classification from one category to the next. If the number of categories from which the rater has to choose is denoted as c, the betadata structure is a c x c matrix with the initial categorization making up the columns and the revised categorization making up the rows.

Using Gwet's notion, these misclassification probabilities are denoted as betas (beta(x|y) where y is the initial category and x is the revised category). For the two category case, this is expresses as a 2 x 2 matrix. For example,

```
beta(1|1) beta(1|2)  = .95   .02
beta(2|1) beta(2|2)    .05   .98
```

indicates that for the generalized rater it is expected that an initial classification of 1 is likely to also be the revised classification 95% of the time, thus leaving the probability of a change of classification at .05 since the sum of the probabilities within each column must sum to 1.00. The likelihood of transition in classification for the second classification level is .02, thus the likelihood for the generalized rater agreeing across classification occasions when originally indicating classification 2 is .98.

This data set would be read into SAS using the following code:

```
data work.betadata;
input b1 b2;
cards;
.95 .02
.05 .98
;
run;
```

The names of the variables are not important. However, for this Beta matrix, the values must be numeric and, as indicated above, the columns must sum to 1.00. Note: to compute the AC1 statistics, the beta matrix of misclassification probabilities is just the identity matrix.

Details on misclassification probabilities:
Suppose a three category situation in which the researcher suspects that raters are likely to be biased to classify objects into level 1. In this case the researchers might posit a misclassification matrix such as:

```
1.0  .4   .4
0    .5   .1
0    .1   .5
```

Here these misclassification probabilities indicate that for the first classification level raters will be certain to agree in their classification on a second occasion. However because of the bias to classify into category 1, the likelihood of changing an initial classification of 2 to a classification of 1 on a second evaluation is considered high (.4) but the likelihood of changing from category 2 to category 3 is low (.1). Similarly, the likelihood of changing an initial classification of 3 to a classification of 1 on a second evaluation is considered high (.4) but the likelihood of change from category 3 to category 2 is low (.1).

****

MACRO OUTPUT:

BETA is the misclassification probabilities provided by the user

ALPHA is the matrix of conditional probabilities of agreement given the observed classification of two categories

Pa is the overall agreement among raters

Pe is the probability of chance agreement

AC is the agreement coefficient estimate, AC1 or AC2.

AgreeStat is the AC1 or AC2 estimate

CVAR is the conditional variance of the AC1 statistic, given the fixed sample of raters. The conditional variance from the current sample of raters without concern for generalizing to the universe of possible raters.

UVAR is the unconditional variance of the AC statistic. The unconditional variance is the estimate of the variance of the AC statistic under the assumption that a subsequent AC would be computed based on a random selection of the same number of raters from all possible raters.

results_data is a temporary SAS dataset that can be saved to a permanent SAS dataset.

results_alpha is a temporary SAS dataset containing the ALPHA matrix for AC2 statistic

```
results_beta is a temporary SAS dataset containing the BETA matrix for AC2 statistic

Note: Local variables "error" and "rc" are created when running this macro.
*******************************************************************************/

%macro AC1AC2(dataset=, numcategories=, betadata=);
title ' '; title2 ' ';
/*-----------------------------------------------------------------------------*/
/*** ERROR Check ***/
%local error rc ;

/*** Check Specification of Beta Matrix, Dataset and Number of Classification Categories ***/
%let error=0;
%if %length(&betadata)=0 %then %do;
        %put ERROR: Beta Matrix must be specified (use of Identity Matrix results in AC1
computation);
        %let error=1;
%end;
%if %length(&dataset)=0 %then %do;
        %put ERROR: Dataset must be specified;
        %let error=1;
%end;
%if %length(&numcategories)=0 %then %do;
        %put ERROR: Number of classification categories must be specified;
        %let error=1;
%end;

/*** Check Beta Matrix Dimension ***/
%let dsid= %sysfunc(open(&betadata));
%if (%sysfunc(attrn(&dsid, NOBS)) ^= %sysfunc(attrn(&dsid, NVARS))) |
        %sysfunc(attrn(&dsid, NVARS))^=&numcategories |
        %sysfunc(attrn(&dsid, NOBS))^=&numcategories %then %do;
        %put ERROR: Beta Matrix must be &numcategories x &numcategories;
        %let error=1;
%end;
%let rc = %sysfunc(close(&dsid));
%if &error=1 %then %GOTO finish;

/*** Check that Beta Matrix Columns Sum to One ***/

%if %length(&betadata) ne 0 %then %do;
        proc means noprint data=&betadata sum;
        output out=_betasums sum=_sum1 - _sum&numcategories;
        run;
        data _null_;
        set _betasums;
        array sums(&numcategories) _sum1 - _sum&numcategories;
        do I= 1 to &numcategories;
                if (sums(I) > 1.0000001) or (sums(I) < 0.9999999) then do;
                        call SYMPUT('error',1);
                end;
        end;
        run;
proc datasets;
delete _betasums;
run; quit;
%end;
%if &error=1 %then %do;
        %put ERROR: Columns of Beta Matrix must sum to 1;
%end;
%if &error=1 %then %GOTO finish;

/*** Check for missing values in Dataset ***/
%if %length(&dataset) ne 0 %then %do;
        data _null_;
        set &dataset;
        array missvals(*) _numeric_;
        if nmiss(of missvals(*)) > 0 then do;
                call SYMPUT ('error', 1);
        end;
        run;
%if &error=1 %then %do;
        %put ERROR: Dataset cannot contain missing values;
```

```
%end;
%if &error=1 %then %GOTO finish;
%end;
/*** End ERROR Check ***/
/*------------------------------------------------------------------------------*/
/*** Begin Computation ***/
proc iml;
print "*** AC1AC2 MACRO OUTPUT ***";
/** Compute both AC1 and AC2 **/
do z = 1 to 2;
/** AC1 **/
if z = 1 then do;
       use &dataset;
       read all into x;
       class_matrix = x;
       use &betadata;
       read all into y;
       beta=I(&numcategories);
end;
/** AC2 **/
if z = 2 then do;
       use &dataset;
       read all into x;
       class_matrix = x;
       use &betadata;
       read all into y;
       beta=y;
end;

/*** Create Dataset Information ***/
q_categories = &numcategories;
n_subjects = nrow(class_matrix);
r_raters = ncol(class_matrix);
if z=1 then do;
       print "NOTE: The number of raters being assumed in this calculation is:" r_raters;
       print "*If this is not correct, check that the first numeric variable in the
              classification dataset contains the ratings from the first rater
              (and not a subject identifier).";
end;
one_vector_n = j(1,n_subjects,1);
one_vector_r = j(r_raters,1,1);
one_vector_q = j(q_categories,1,1);

/*** Create Agreement Matrix ***/
do i = 1 to q_categories;
       if i = 1 then do;
       agreeMat = (class_matrix=1)*one_vector_r;
       end;
       else do;
       agreeMat = agreeMat||((class_matrix=i)*one_vector_r);
       end;
end;

/***** AC STATISTIC ****/
/*** Create PiQ ***/
/* t(nxq) t(1xn) -> qx1 */
pi_lvec = (agreeMat` * one_vector_n`) / (n_subjects # r_raters);
/* qxq qx1 -> qx1 */
pi_qvec = beta * pi_lvec;

/*** Create Pe ***/
/*  t(qx1) qx1 -> 1x1 */
p_e = one_vector_q` * ((pi_qvec # (1-pi_qvec))/(q_categories-1));


/*** Create Alpha Matrix ***/
alpha = j(q_categories, q_categories, 0); /* create qxq 0 matrix */
do i = 1 to q_categories;
       do j = 1 to q_categories;
              /* qx1 qxq -> qxq */
              alpha[i,j] = one_vector_q` * (beta[,i]#beta[,j]);
       end;
end;
```

```
/** Create Pa ***/
paVec = j(n_subjects,1,0);
do i = 1 to n_subjects;
       rr = 0 # beta; /* creates qxq matrix of 0's */
       do j = 1 to q_categories;
              do l = 1 to q_categories;
                     if j=l then do;
                            rr[j,l] = alpha[j,l] # agreeMat[i,j] # (agreeMat[i,j]-1);
                     end;
                     if j ^= l then do;
                            rr[j,l] = alpha[j,l] # agreeMat[i,j] # agreeMat[i,l];
                     end;
              end; /* end of l */
       end; /* end of j */
       /* t(qx1) qxq qx1 -> 1x1 */
       paVec[i,1] = one_vector_q` * rr * one_vector_q;
end; /* end of n */

/* 1xn nx1 -> 1x1 */
p_a = (one_vector_n * paVec)/ (n_subjects # r_raters # (r_raters-1));

/*** Create AC Statistic ***/
delta_second = (p_a - p_e)/(1-p_e);
AC_statistic = delta_second;


/*** Estimate Variance of AC ***/
/** Need individual components of pa for variance estimation **/
/* nx1 */
pa_iVec = paVec/(r_raters#(r_raters-1));
k = (pa_iVec - p_e)/(1-p_e);
/* 1xn nx1 -> 1x1 */
s2 = (one_vector_n * (k - delta_second)##2) / (n_subjects-1);
/***** CONDITIONAL VARIANCE OF AC *****/
cvar = s2/n_subjects; * assuming negligeable sampling fraction, f=n/N *;

/****** UNCONDITIONAL VARIANCE OF AC ****/
m1_matrix = j(n_subjects, n_subjects,0); ** nxn matrix **;
do q1 = 1 to q_categories;
       do q2 = 1 to q_categories;
       mq1q2_matrix = (class_matrix=q1) * ((class_matrix=q2)`);
       m1_matrix = m1_matrix + alpha[q1,q1]#alpha[q2,q2]#(mq1q2_matrix # (mq1q2_matrix-1));
       end; /* end of q1 */
end; /* end of q2 */

m2_matrix = j(n_subjects, n_subjects,0); /** nxn matrix **/
do h1 = 1 to q_categories;
       do h2 = 1 to q_categories;
              do q1 = 1 to q_categories;
                     do q2 = 1 to q_categories;
                            if q1 ^= h1 & q2 ^= h2 then do;
                                   mq1q2_matrix = (class_matrix=q1) * ((class_matrix=q2)`);
                                   mh1h2_matrix = (class_matrix=h1) * ((class_matrix=h2)`);
                                   m2_matrix = m2_matrix + (alpha[q1,h1]# alpha[q2,h2]#
mq1q2_matrix # (mh1h2_matrix));
                            end;
                            if q1 = h1 | q2 = h2 then do;
                                   m2_matrix = m2_matrix;
                            end;
                     end; /* end of q1 */
              end; /* end of q2 */
       end; /* end of h2 */
end; /* end of h1 */

m_matrix = (m1_matrix + m2_matrix) / (r_raters#(r_raters - 1));
 /* 1xn nxn t(1xn) */
p_2a = (one_vector_n * m_matrix * one_vector_n`) / (n_subjects##2);

/*** Need to Compute M Matrix for each pair of subjects ***/
/** Create P_app ***/
pappVec = j(n_subjects,1,0);
do i = 1 to n_subjects;
       rr_p = j(q_categories,q_categories,0); /* creates qxq matrix of 0's */
       do j = 1 to q_categories;
```

```
                do l = 1 to q_categories;
                        if j=l then do;
                                rr_p[j,l] = (alpha[j,l]##2) # agreeMat[i,j] # (agreeMat[i,j]-1);
                        end;
                        if j ^= l then do;
                                rr_p[j,l] = (alpha[j,l]##2) # agreeMat[i,j] # agreeMat[i,l];
                        end;
                end; /* end of l */
        end; /* end of j */
        /* t(qx1) qxq qx1 -> 1x1 */
        pappVec[i,1] = one_vector_q` * rr_p * one_vector_q;
end; /* end of n */

/* 1xn nx1 -> 1x1 */
p_app = (one_vector_n * pappVec)/ (n_subjects # r_raters # (r_raters-1));

cr = 1/(r_raters#(r_raters-1)#((1-p_e)##2));
 /** Unconditional Variance of AC2 assuming negligible sampling fraction f=n/N **/
uvar = ( s2/n_subjects + cr#(p_2a + (p_app - p_2a)/n_subjects) );

/*************************************************************************/
/*** Print Selected Output ***/
if z=1 then do;
print "AC1 Alpha and Beta Matrix";
end;
if z=2 then do;
print "AC2 Alpha and Beta Matrix";
print "(AC2 Alpha Matrix Saved in Temporary SAS Dataset, results_alpha_AC2)";
end;
print "beta: Transition Probabilities" beta;
print "alpha: Conditional Agreement Probabilities" alpha;

/*************************************************************************/
/** OUTPUT RESULTS TO SAS DATASETS **/
/* Statistics */
if z=1 then do;
        results = j(2,6,0);
        results[z,1] = 1;
end;
if z=2 then do;
        results = results;
        results[z,1] = 2;
end;
results[z,2] = p_a;
results[z,3] = p_e;
results[z,4] = delta_second;
results[z,5] = cvar;
results[z,6] = uvar;

/* Beta Matrix for AC2 */
if z=2 then do;
        create results_beta_AC2 from beta;
        append from beta;
end;
/* Alpha Matrix for AC2 */
if z=2 then do;
        create results_alpha_AC2 from alpha;
        append from alpha;
end;
end; /* end of z (AC1 vs. AC2) indicator */

create results_data from results[colname={"AC","PA","PE","AgreeStat","CondVar","UncondVar"}];
append from results;
quit;

/*** Print Output Dataset ***/
proc print data=results_data noobs;
var AC AgreeStat CondVar UncondVar PA PE;
title 'AC1 and AC2 Information';
title2 '(Saved in Temporary SAS Dataset, results_data)';
run;
title; title2;
%finish:
%mend;
```