Series: Statistical Methods For Inter-Rater Reliability Assessment, No. **2**, May 2002

Inter-Rater Reliability: Dependency on Trait Prevalence and Marginal Homogeneity

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

Abstract. Researchers have criticized chance-corrected agreement statistics, particularly the Kappa statistic, as being very sensitive to raters' classification probabilities (marginal probabilities) and to trait prevalence in the subject population. Consequently, several authors have suggested that marginal probabilities be tested for homogeneity and that any comparison between reliability studies be preceded by an assessment of trait prevalence among subjects. The objective of this paper is threefold: (*i*) to demonstrate that marginal homogeneity testing does not prevent the unpredictable results often obtained with some of the most popular agreement statistics, (*ii*) to present a simple and reliable inter-rater agreement statistic, and (*iii*) to gain further insight into the dependency of agreement statistics upon trait prevalence.

1. Introduction

To study the dependency of Kappa statistic on marginal probabilities and trait prevalence, let us consider a reliability study where two observers **A** and **B** must classify each of the n study subjects as positive (**+**) if they carry a specific trait of interest, or as negative (**−**) otherwise. The outcome of this experiment is described in Table 1.

Table 1: Distribution of n subjects by rater and response

Rater B	Rater A		Total
	$+$	$-$	
$+$	a	b	B_+
$-$	c	d	B_-
Total	A_+	A_-	n

Table 1 indicates that both raters have classified a of the n subjects into category "+" and d subjects into category "−". There are b subjects that raters A and B have classified into categories "−" and "+" respectively. Similarly, the same c subjects that rater A classified in category "+" were also classified in category "−" by rater B. The total number of subjects that rater A classified into the positive and negative categories are respectively denoted by A_+ and A_- (i.e. $A_+ = a + c$ and $A_- = b + d$). The marginal totals B_+ and B_- can be defined in a similar way.

The 3 most popular inter-rater agreement statistics used to quantify the extent of agreement between raters A and B on the basis of agreement data of Table 1 are the S coefficient of Bennet et al. (1954), the π-statistic (read PI-statistic) suggested by Scott (1955) and the κ-statistic (read KAPPA-statistic) of Cohen (1960). Each of these statistics is a function of the overall agreement probability P_a, which for all practical purposes, is the proportion of subjects that both raters classified into the same categories. We have that,

$$P_a = \frac{a + d}{n}. \qquad (1)$$

The inter-rater reliability as measured by the $\mathcal{S}$-coefficient is defined as follows:

$$\mathcal{S} = 2P_a - 1\cdot \tag{2}$$

Scott's π-statistic is given by:

$$\mathcal{PI} = \frac{P_a - P_e(\pi)}{1 - P_e(\pi)}, \tag{3}$$

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

$$P_e(\pi) = P_+^2 + P_-^2, \tag{4}$$

where $P_+ = \big[(A_+ + B_+)/2\big]/n$ and $P_- = \big[(A_- + B_-)/2\big]/n$. Some authors have mistakenly criticized Scott's π-statistic by claiming it was developed under the assumption of marginal homogeneity. The quantity P_+ should not be interpreted as a common probability that each rater classify a subject into the positive category. It should rather be interpreted as the probability that a randomly chosen rater (A or B) classify a randomly chosen subject into category "+". The probability P_- can be interpreted in a similar way. This interpretation, also suggested by Fleiss (1971) seems to be consistently misunderstood.

Cohen's κ-statistic is given by:

$$\mathcal{KA} = \frac{P_a - P_e(\kappa)}{1 - P_e(\kappa)}, \tag{5}$$

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:

$$P_e(\kappa) = P_{A+}P_{B+} + P_{A-}P_{B-}, \tag{6}$$

where $P_{A+} = A_+/n$, $P_{A-} = A_-/n$, $P_{B+} = B_+/n$, and $P_{B-} = B_-/n$. Cohen's Kappa statistic uses rater-specific classification probabilities P_{A+}, P_{A-}, P_{B+}, and P_{B+} to compute the likelihood of agreement by chance, while Scott's approach is based on the overall classification probabilities P_+ and P_-.

2. Dependency of Agreement Measures on Bias and Prevalence

The 2 basic questions one should ask about any inter-rater reliability statistic are the following:

- Does the statistic provide a measure of the extent of agreement between raters that is easily interpretable?
- How precise are the methods for obtaining them?

We believe that while the limitations of agreement statistics presented in the literature are justified and should be addressed, many researchers have had a tendency of placing unduly high expectations upon chance-corrected agreement measures. These are summary measures, intended to provide a broad picture of the agreement level between raters and cannot be used to answer each specific question about the reliability study. Agreement statistics can and should be suitably modified to address specific concerns. Cicchetti and Feinstein (1990) pointed out the fact that a researcher may want to know what was found for agreement in positive and negative responses. This is a specific problem that could be resolved by conditioning any agreement statistic on the specific category of interest (see Gwet (2001), chapter 8).

The PI and Kappa statistics are two chance-corrected agreement measures that are highly sensitive to the trait prevalence in the subject population as well as to differences in rater marginal probabilities P_{A+}, P_{A-}, P_{B+}, and P_{B-}. These 2 properties make the PI and Kappa statistics very unstable and often difficult to interpret. The $\mathcal{S}$ coefficient is also difficult to interpret for different reasons that are discussed in subsequent sections. This section contains a few examples of inter-rater reliability experiments where interpreting these statistics is difficult.

Let us consider a researcher who conducted 4 inter-rater reliability experiments (experiments $\boldsymbol{E1}$, $\boldsymbol{E2}$, $\boldsymbol{E3}$, $\boldsymbol{E4}$) and reported the outcomes respectively in Tables 2, 3, 4, and 5.

Marginal totals in Table 2 suggest that raters $\boldsymbol{A}$ and $\boldsymbol{B}$ tend to classify subjects into categories with similar frequencies. Marginal totals in Table 3 on the other hand, indicate that rater $\boldsymbol{B}$ is much more likely to classify a subject into the positive category than rater $\boldsymbol{A}$, who in turn (a fortiori) is more likely to classify a subject into the negative category. *Most researchers would prefer in such situations, an agreement statistic that would yield a higher inter-rater reliability for experiment* $\boldsymbol{E1}$ *than for experiment* $\boldsymbol{E2}$. By favoring the opposite conclusion, neither Scott's PI statistic nor Cohen's Kappa statistic are consistent with this researchers' expectation. The $\boldsymbol{\mathcal{S}}$ coefficient gives the same inter-rater reliability for both experiments and does not satisfy that requirement either. Note that raters $\boldsymbol{A}$ and $\boldsymbol{B}$ in experiments $\boldsymbol{E1}$ and $\boldsymbol{E2}$ have the same overall agreement probability $\boldsymbol{P_a} = (45+15)/100 = (25+35)/100 = 0.6$.

- For experiment $\boldsymbol{E1}$ (Table 2), we have that $\boldsymbol{P_+} = [(60+70)/2]/100 = 0.65$ and $\boldsymbol{P_-} = [(40+30)/2]/100 = 0.35$, which leads to a chance-agreement probability $\boldsymbol{P_e(\pi)} = (0.65)^{\boldsymbol{2}} + (0.35)^{\boldsymbol{2}} = 0.545$, according to Scott's method. Therefore, the PI statistic is given by: $\boldsymbol{\mathcal{PI}} = (0.6 - 0.545)/(1 - 0.545) = 0.1209$.

  To obtain the Kappa statistic, one would notice that chance-agreement probability according Cohen's method is given by $\boldsymbol{P_e(\kappa)} = 0.60 \times 0.70 + 0.40 \times 0.30 = 0.54$. This leads to a Kappa statistic $\boldsymbol{\mathcal{KA}} = (0.60 - 0.54)/(1 - 0.54) = 0.1304$.

- For experiment $\boldsymbol{E2}$ (Table 3), we have that $\boldsymbol{P_+} = [(60+30)/2]/100 = 0.45$ and $\boldsymbol{P_-} = [(40+70)/2]/100 = 0.55$, which leads to Scott's chance-agreement probability of $\boldsymbol{P_e(\pi)} = (0.45)^{\boldsymbol{2}} + (0.55)^{\boldsymbol{2}} = 0.505$. Therefore, the PI statistic is given by: $\boldsymbol{PI} = (0.6 - 0.505)/(1 - 0.505) = 0.1920$.

  For this experiment ($\boldsymbol{E2}$), chance-agreement probability according to Cohen's proposal is $\boldsymbol{P_e(\kappa)} = 0.60 \times 0.30 + 0.40 \times 0.70 = 0.46$. This leads to a Kappa statistic of $\boldsymbol{\mathcal{KA}} = (0.60 - 0.46)/(1 - 0.46) = 0.2593$.

While most researchers would expect the extent of agreement between raters $\boldsymbol{A}$ and $\boldsymbol{B}$ to be higher for experiment $\boldsymbol{E1}$ than for experiment $\boldsymbol{E2}$, the 2 most popular inter-rater agreement statistics reach a different conclusion. This unwanted dependency of chance-corrected statistics on the differences in raters' marginal probabilities is problematic and must be resolved to reestablish the credibility of the idea of correcting for agreement by chance. This problem is formally studied in section 3 where we make some recommendations.

Table 2: Distribution of 100 subjects by rater and response: experiment $\boldsymbol{E1}$

Rater $\boldsymbol{B}$	Rater $\boldsymbol{A}$		Total
	+	−	
+	**45**	**15**	**60**
−	**25**	**15**	**40**
Total	**70**	**30**	**100**

Table 3: Distribution of 100 subjects by rater and response: experiment $\boldsymbol{E2}$

Rater $\boldsymbol{B}$	Rater $\boldsymbol{A}$		Total
	+	−	
+	**25**	**35**	**60**
−	**5**	**35**	**40**
Total	**30**	**70**	**100**

For experiments $\boldsymbol{E3}$ and $\boldsymbol{E4}$ (Tables 4 and 5), the raters have identical marginal probabilities. Table 4 indicates that raters $\boldsymbol{A}$ and $\boldsymbol{B}$ in experiment $\boldsymbol{E3}$ will both classify a subject into either

category ("+" or "−") with the same probability of 50%. Similarly, it follows from Table 5 that raters $\boldsymbol{A}$ and $\boldsymbol{B}$ in experiment $\boldsymbol{E4}$ would classify a subject into the positive category with the same probability of 80% and into the negative category with the same probability of 20%. Because the raters have the same propensity for positive (and a fortiori negative) rating in these 2 experiments, researchers would expect them to have high inter-rater reliability coefficients in both situations.

The Kappa and PI statistics have failed again to meet researchers' expectations in the situations described in Tables 4 and 5, where the overall agreement probability $\boldsymbol{P_a}$ is evaluated at 0.8.

- For experiment $\boldsymbol{E3}$ (Table 4), we have $\boldsymbol{P_+} = \boldsymbol{P_-} = [(50 + 50)/2]/100 = 0.50$. This leads to a Scott's chance-agreement probability of $\boldsymbol{P_e(\pi)} = (0.5)^{\boldsymbol{2}} + (0.5)^{\boldsymbol{2}} = 0.5$. Thus, the PI statistic is given by: $\boldsymbol{\mathcal{PI}} = (0.8 - 0.5)/(1 - 0.5) = 0.60$.

  The Kappa statistic is obtained by first calculating Cohen's chance-agreement probability $\boldsymbol{P_e(\kappa)}$ as follows: $\boldsymbol{P_e(\kappa)} = 0.5 \times 0.5 + 0.5 \times 0.5 = 0.5$. This leads to a Kappa statistic that is identical to the PI statistic obtained in the previous paragraph (i.e. $\boldsymbol{\mathcal{KA}} = 0.60$).

- For experiment $\boldsymbol{E4}$ (Table 5), we have that $\boldsymbol{P_+} = [(80 + 80)/2]/100 = 0.80$ and $\boldsymbol{P_-} = [(20 + 20)/2]/100 = 0.20$, which leads to Scott's chance-agreement probability of $\boldsymbol{P_e(\pi)} = (0.8)^{\boldsymbol{2}} + (0.2)^{\boldsymbol{2}} = 0.68$. Thus, the PI statistic is given by: $\boldsymbol{\mathcal{PI}} = (0.8 - 0.68)/(1 - 0.68) = 0.375$.

  Using Cohen's approach, chance-agreement probability is obtained as follows: $\boldsymbol{P_e(\kappa)} = 0.8 \times 0.8 + 0.2 \times 0.2 = 0.68$. Again, the Kappa statistic is identical to the PI statistic and given by $\boldsymbol{\mathcal{KA}} = 0.375$.

It appears from Tables 4 and 5 that even when the raters share the same marginal probabilities, the prevalence of the trait in the population may dramatically affect the magnitude of the inter-rater reliability as measured by the Kappa and PI statistics. In fact, there is no apparent reason why Tables 4 and 5 would not both yield high agreement coefficients. One would expect a lower chance-agreement probability from Table 5, as both raters appear to be more inclined to classify subjects into the positive category, leading to an agreement coefficient that is higher for Table 5 than for Table 4.

Table 4: Distribution of 100 subjects by rater and response: experiment $\boldsymbol{E3}$

Rater $\boldsymbol{B}$	Rater $\boldsymbol{A}$		Total
	+	−	
+	**40**	**10**	**50**
−	**10**	**40**	**50**
Total	**50**	**50**	**100**

Table 5: Distribution of 100 subjects by rater and response: experiment $\boldsymbol{E4}$

Rater $\boldsymbol{B}$	Rater $\boldsymbol{A}$		Total
	+	−	
+	**70**	**10**	**80**
−	**10**	**10**	**20**
Total	**80**	**20**	**100**

Until now, we have been focussing on the deficiencies of Kappa and PI statistic, and we have said very little about the performance of the $\boldsymbol{\mathcal{S}}$ coefficient. Although the $\boldsymbol{\mathcal{S}}$ coefficient is not affected by differences in marginal probabilities, nor by the trait prevalence in the population under study, it suffers from another equally serious problem. In fact, the extent of agreement between 2 raters obtained with the $\boldsymbol{\mathcal{S}}$ coefficient will always be 0 or close to 0 if both raters agree about 50% of the times, and will be negative whenever they classify less than 50% of subjects into the same category.

This is a major limitation since the correct classification of 50% of subjects should indicate some agreement level between the 2 raters.

The examples presented in this section were necessary to point out some unexpected results that one may get when the inter-rater agreement measures proposed in the literature are used. The next section is devoted to a deeper analysis of these problems.

3. Sensitivity Analysis and Alternative Agreement Measure

In section 2, we have discussed about the limitations of the most popular inter-rater agreement statistics proposed in the literature. The goal in this section is to show how the $\mathcal{AC}_1$ statistic introduced by Gwet (2002) resolves all the problems discussed in section 2, and to conduct a formal analysis of the sensitivity of agreement indices to trait prevalence and marginal differences.

The $\mathcal{AC}_1$ statistic is given by:

$$\mathcal{AC}_1 = \frac{P_a - P_e(\gamma)}{1 - P_{e(\gamma)}}, \qquad (7)$$

where $P_e(\gamma)$ is defined as follows:

$$P_e(\gamma) = 2P_+(1 - P_+). \qquad (8)$$

We believe that the $\mathcal{AC}_1$ statistic overcomes all the limitations of existing agreement statistics, and should provide a valuable and more reliable tool for evaluating the extent of agreement between raters. It was extended to more general multiple-rater and multiple-item response reliability studies by Gwet (2002), where precision measures and statistical significance are also discussed.

For the situations described in Tables 2 and 3, it is easy to verify that $P_e(\gamma)$ takes respectively the values of 0.455 and 0.495, leading to an $\mathcal{AC}_1$ statistic of 0.27 and 0.21. As one would expect from the marginal differences, the outcome of experiment $E1$ yields a higher inter-rater reliability than the outcome of experiment $E2$.

As mentioned earlier, Tables 4 and 5 describe 2 situations where researchers would expect high agreement coefficients. It can be verified that $P_e(\gamma) = 0.5$ for Table 4 and $P_e(\gamma) = 0.32$ for Table 5. This leads to an $\mathcal{AC}_1$ statistic of 0.6 and 0.71 for Tables 4 and 5 respectively. Although both Tables show the same overall agreement probability $P_a = 0.80$, the high concentration of observations in one cell of Table 5 has led to a smaller chance-agreement probability for that Table. Thus, the resulting agreement coefficient of 0.71 is higher than that of Table 4. Unlike the Kappa and PI statistics, the $\mathcal{AC}_1$ statistic provides results that are consistent with what researchers expect.

3.1 Sensitivity to Marginal Homogeneity

To study the sensitivity of Kappa, PI, and $\mathcal{AC}_1$ statistics to the differences in marginal probabilities, we have conducted an experiment where the overall agreement probability P_a is fixed at 0.80, and the value a of the (+,+) cell (see Table 1) varies from 0 to 0.80 by increment of 0.10. The results are depicted in figure 1, where the variation of the 3 agreement statistics is graphically represented as a function of the overall probability P_+ ($P_+ = (A_+ + B_+)/2$) of classification into the positive category.

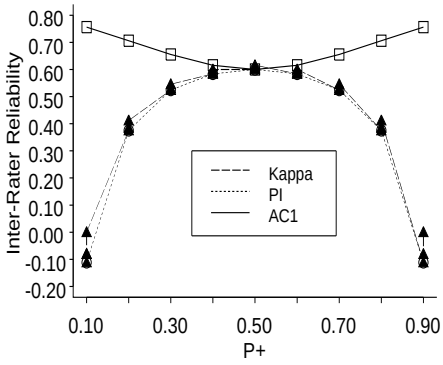

Figure 1: Kappa, PI, and $\mathcal{AC}_1$ statistics as a function of P_+, when $P_a = 0.80$.

Note that for a given value of P_a, there is a one-to-one relationship between P_+ and the $\mathcal{PI}$ statistic on one hand, and between P_+ and

the $\mathcal{AC}_1$ statistic on the other hand. However, each value of P_+ is associated with 5 different pairs (A_+, B_+) that will yield 3 values for the Kappa statistic. This explains the presence on the graph of 3 points ▲ corresponding to the kappa values associated with each value of P_+. For example, when $P_+ = 0.1$, the pair (A_+, B_+) will take one of the values $\{(0.0, 0.20), (0.05, 0.15), (0.10, 0.10), (0.15, 0.05), (0.20, 0.00)\}$, leading to Kappa values of 0, -0.0811,-0.1111,-0.0811, and 0 respectively.

Several remarks can be made about figure 1.

- Figure 1 indicates that when the overall propensity of positive rating P_+ is either very small or very large, the Kappa and PI statistics tend to yield dramatically low agreement coefficients, wether marginal probabilities are identical or not. Therefore, the commonly-accepted argument according to which agreement data should be tested for marginal homogeneity before using chance-corrected agreement statistics is misleading. In fact, marginal homogeneity does not guarantee the validity of the results.
- All 3 agreement statistics represented in figure 1 take a common value of 0.60 only when $A_+ = B_+ = 0.5$. When both marginal probabilities are not identically equal to 0.5, but are close enough to 0.5, all 3 statistics will still yield very similar agreement levels.
- In this experiment, the agreement probability P_a was fixed at 0.80, which is expected to yield a high inter-rater reliability. Only the $\mathcal{AC}_1$ statistic consistently gives a high agreement coefficient. The $\mathcal{S}$ coefficient was not used in this experiment as it takes the same value of 0.6 for all P_+. A very low or very high value for P_+ is characterized by a high concentration of observation in one table cell. This should reduce the magnitude of chance-agreement probability, leading to a higher inter-rater reliability. Only the $\mathcal{AC}_1$ statistic seems to capture this phenomenon.

For comparability, we have changed the value of P_a from 0.8 to 0.6 so that we can observe how the graph in figure 1 would be affected. The results shown in figure 2, lead to the same conclusions as in figure 1.

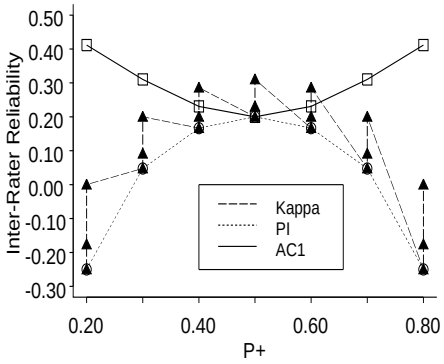

Figure 2: Kappa, PI, and $\mathcal{AC}_1$ statistics as a function of P_+, when $P_a = 0.60$.

3.2. Sensitivity to Trait Prevalence

To evaluate how sensitive the $\mathcal{AC}_1$, $\mathcal{S}$, $\mathcal{KA}$, and $\mathcal{PI}$ statistics are to the prevalence of a trait in the population, it is necessary to introduce 2 concepts often used in epidemiology: the *sensitivity* and the *specificity* of a rater. These concepts are introduced under the assumption that there exists a "gold standard" that provides the "correct" classification of subjects.

The **sensitivity** associated with a rater A and denoted by α_A, is the conditional probability that the rater correctly classify a subject into the positive category. It represents the propensity of a rater to identify positive cases.

The **specificity** associated with a rater A and denoted by β_A, is the conditional probability that the rater correctly classify a subject into the negative category. It represents the propensity of a rater to identify negative cases.

The **prevalence** of trait in a subject population, denoted by P_r, is the probability that a randomly chosen subject from the population turn

out to be positive. For all practical purposes, it represents the proportion of positive cases in the subject population.

To study how trait prevalence affects agreement statistics, we must assign a constant value to each rater's sensitivity and specificity so as to obtain the expected extent of agreement between raters. Let P_{A+} and P_{B+} denote respectively the probabilities that raters A and B classify a subject as positive. Using standard probabilistic considerations, these 2 probabilities can be expressed as follows:

$$P_{A+} = P_r\alpha_A + (1-P_r)(1-\beta_A), \quad (9)$$

$$P_{B+} = P_r\alpha_B + (1-P_r)(1-\beta_B). \quad (10)$$

For given values of P_r, α_A, α_B, β_A and β_B, the expected marginal totals A_+ and B_+ of the positive category (see Table 1) are obtained as, $A_+ = nP_{A+}$ and $B_+ = nP_{B+}$. Moreover, raters A and B will both classify a subject as positive with a probability P_{++} that is given by:

$$P_{++} = P_r\alpha_A\alpha_B + (1-P_r)(1-\beta_A)(1-\beta_B) \quad (11)$$

The probability for both raters to classify a subject into the negative category ("−") is given by:

$$P_{--} = 1 - \Big(P_{A+} + P_{B+} - P_{++}\Big). \quad (12)$$

Since the overall agreement probability P_a that goes into the calculation of all agreement statistics is the sum of P_{++} and P_{--}, we can compute the Kappa, PI, $\mathcal{S}$ and $\mathcal{AC}_1$ statistics as functions of prevalence (P_r) using equations 9, 10, 11, and 12 for preassigned values of raters' sensitivity and specificity.

Some researchers see the dependence of chance-corrected agreement statistics on trait prevalence as a drawback. We believe that it is not. What is questionable, is the nature of the relationship between $\mathcal{KA}$ or $\mathcal{PI}$ statistics and trait prevalence. A good agreement measure should have certain properties that are suggested by common sense. For example,

1. If raters' sensitivity and specificity are all equal and are high, then the inter-rater reliability should be high when the trait prevalence is either low or high.
2. If raters' sensitivity is smaller than their specificity, then one would expect a higher inter-rater reliability when the trait prevalence is small.
3. If raters' sensitivity is greater than their specificity, then one would expect a higher inter-rater reliability when the trait prevalence is high.

One should note that a high sensitivity indicates that the raters easily detect positive cases. Therefore, high sensitivity together with high prevalence should lead to high inter-rater reliability. This explains why the inter-rater reliability should have the 3 properties stated above to be easily interpretable. Unfortunately, Kappa and $\mathcal{PI}$ statistics do not have any of these properties. The $\mathcal{S}$ coefficient have these properties but will usually underestimate the extent of agreement between the raters, while the $\mathcal{AC}_1$ statistics have all these properties in addition to providing the right correction for chance agreement.

To illustrate the dependency of the various agreement statistics upon prevalence, we have preassigned a constant value of 0.9 to raters' sensitivity and specificity and let the prevalence rater P_r vary from 0 to 1 by increment of 0.10. The agreement statistics were obtained using equations 9, 10, 11, 12, and the results reported in Table 6. These results are depicted in figure 3, where the inter-rater reliability coefficients as measured by P_a, $\mathcal{S}$, $\mathcal{KA}$, $\mathcal{PI}$, and $\mathcal{AC}_1$, are plotted against trait prevalence.

It follows from Table 6 that when the raters' sensitivity and specificity are equal to 0.9, then the expected overall agreement probability P_a will take a unique value of 0.82. Kappa and PI statistics take identical values that vary from 0 to 0 with a maximum of 0.64 when the prevalence rate is at 0.5. The behavior of Kappa and PI statistics is difficult to explain. In fact, with

a prevalence of 100% and high raters' sensitivity and specificity, Kappa and PI still produce an inter-rater reliability estimated at 0.

Table 6: Inter-Rater Agreement as a Function of Prevalence when $\alpha_A = \alpha_B = \beta_A = \beta_B = 0.9$

P_r	P_a	$\mathcal{KA}$	$\mathcal{PI}$	$\mathcal{AC}_1$	$\mathcal{S}$
0.00	0.82	0.00	0.00	0.78	0.64
0.01	0.82	0.07	0.07	0.78	0.64
0.05	0.82	0.25	0.25	0.76	0.64
0.10	0.82	0.39	0.39	0.74	0.64
0.20	0.82	0.53	0.53	0.71	0.64
0.30	0.82	0.60	0.60	0.67	0.64
0.40	0.82	0.63	0.63	0.65	0.64
0.50	0.82	0.64	0.64	0.64	0.64
0.60	0.82	0.63	0.63	0.65	0.64
0.70	0.82	0.60	0.60	0.67	0.64
0.80	0.82	0.53	0.53	0.71	0.64
0.90	0.82	0.39	0.39	0.74	0.64
0.95	0.82	0.25	0.25	0.76	0.64
0.99	0.82	0.07	0.07	0.78	0.64
1.00	0.82	0.00	0.00	0.78	0.64

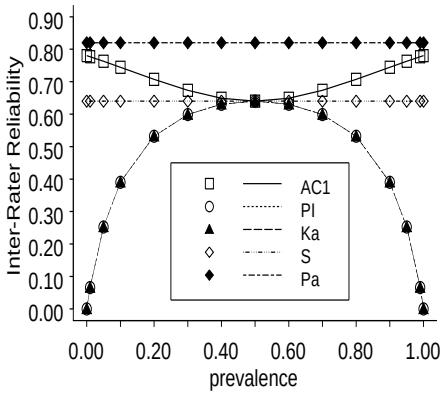

Figure 3: P_a, $\mathcal{S}$, Kappa, PI, and $\mathcal{AC}_1$ statistics as a function of P_r, when $\alpha_A = \alpha_B = \beta_A = \beta_B = 0.90$.

The $\mathcal{S}$ coefficient's curve is parallel to that of P_a as it applies a uniform maximum chance-agreement correction of 0.5. The $\mathcal{AC}_1$ will usually take a value between P_a and $\mathcal{S}$ after applying an optimum correction for chance agreement.

In Table 7, we assumed that the raters have a common sensitivity of 0.8 and a common specificity of 0.9. Kappa and PI statistics vary again from 0 to 0 with a maximum of 0.50 reached at a prevalence rate of 0.4. For reasons discussed in the last paragraph, this behavior is not consistent with any expectations based on common sense. The scattered plot corresponding to Table 7 is shown in figure 4, where the $\mathcal{AC}_1$ line is located between that of the agreement probability P_a and that of the $\mathcal{S}$ coefficient.

Table 7: Inter-Rater Agreement as a Function of Prevalence when $\alpha_A = \alpha_B = 0.8$ and $\beta_A = \beta_B = 0.9$

P_r	P_a	$\mathcal{KA}$	$\mathcal{PI}$	$\mathcal{AC}_1$	$\mathcal{S}$
0.00	0.82	0.00	0.00	0.78	0.64
0.01	0.82	0.05	0.05	0.78	0.64
0.05	0.81	0.20	0.20	0.76	0.63
0.10	0.81	0.31	0.31	0.73	0.61
0.20	0.79	0.43	0.43	0.67	0.58
0.30	0.78	0.48	0.48	0.61	0.56
0.40	0.76	0.50	0.50	0.55	0.53
0.50	0.75	0.49	0.49	0.50	0.50
0.60	0.74	0.47	0.47	0.47	0.47
0.70	0.72	0.43	0.43	0.46	0.44
0.80	0.71	0.35	0.35	0.47	0.42
0.90	0.69	0.22	0.22	0.49	0.39
0.95	0.69	0.13	0.13	0.51	0.37
0.99	0.68	0.03	0.03	0.53	0.36
1.00	0.68	0.00	0.00	0.53	0.36

The fact that the overall agreement probability P_a decreases as the prevalence rate P_r increases is interesting. Since P_a is the primary ingredient going into the calculation of the inter-rater reliability coefficient, it is expected that any measure of the extent of agreement between raters will be affected by the trait prevalence. However, a chance-corrected measure should remain reasonably close to the quantity that it is supposed to adjust for chance agreement. As shown in figure 4, $\mathcal{KA}$ and $\mathcal{PI}$ statistics are sometimes dramati-

cally low and much smaller than P_a.

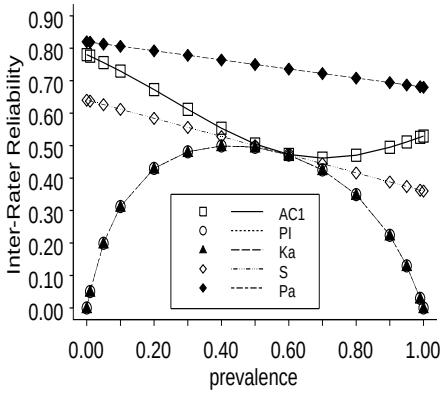

Figure 4: P_a, $\mathcal{S}$, Kappa, PI, and $\mathcal{AC}_1$ statistics as a function of P_r, when $\alpha_A = \alpha_B = 0.80$ and $\beta_A = \beta_B = 0.90$

4. Concluding Remarks

Behavioral scientists and other researchers have always been concerned about the dependency of chance-corrected agreement statistics on trait prevalence in the subject population as well as on marginal homogeneity. In this article, we have conducted a sensitivity analysis to gain further insight into these issues. We have demonstrated that the widely-used Kappa and PI statistics have a suspicious behavior with respect to the variation of the trait prevalence rate and to the magnitude of raters' classification probabilities. More specifically we have established the following:

- Kappa and PI statistics are more affected by the overall propensity of positive classification ($P_+ = (A_+ + B_+)/2$) than by the differences in marginal probabilities. Therefore, marginal homogeneity testing is not necessarily a relevant practice.
- $\mathcal{PI}$ and $\mathcal{KA}$ statistics yield reasonable results when P_+ is close to 0.5. In other words if raters A and B classify a subject into the "+" category with probabilities that sum to 1, then the inter-rater reliability obtained from the $\mathcal{PI}$ and $\mathcal{KA}$ statistics will be reasonable.
- All agreement statistics depend on the magnitude of the trait prevalence rate. However, the variation of $\mathcal{KA}$ and $\mathcal{PI}$ statistics as functions of trait prevalence is very erratic. These 2 statistics would yield very low agreement coefficients for low or high prevalence when any researcher would expect a very high inter-rater reliability.
- The $\mathcal{AC}_1$ statistic seems to have the best statistical properties of all agreement statistics discussed in this paper.

5. 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

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

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

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

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
Phone: 301-947-6652, Fax: 301-947-2959.