Kappa Statistic is not Satisfactory for Assessing the Extent of Agreement Between Raters

(April 2002)

Kilem Gwet, Ph.D. (kilem62@yahoo.com)
Sr. Statistical Consultant, STATAXIS Consulting
15914B Shady Grove Road, PMB 145, Gaithersburg, MD 20877. U.S.A.

1. ABSTRACT

Evaluating the extent of agreement between 2 or between several raters is common in social, behavioral and medical sciences. The objective of this paper is to provide a detailed discussion about the limitations of the kappa statistic, which is a commonly used technique for computing the inter-rater reliability coefficient.

2. INTRODUCTION

Two statistics are often used in practice for evaluating the extent of agreement between raters. These are the π-statistic (should be read "pi-statistic") suggested by Scott (1955) and the Kappa statistic suggested by Cohen (1960). Researchers often use the π-statistic and mistakenly refer to it as Kappa. It is necessary to distinguish these 2 statistics, which will produce similar results most of the times. This will especially be true when the agreement rate between raters is reasonably high.

After the π-statistic and the Kappa statistic are defined, a simple example will be presented to show the inadequacy of Kappa for evaluating the extent of agreement between 2 raters. This example is followed by a discussion about the causes of the problem. Afterwards, a simple alternative chance-corrected statistic that consistently yields reliable results will be recommended.

Let us consider a reliability experiment where 2 raters (or observers, or judges) referred to as **rater A** and **rater B** must classify N subjects into one of 2 possible response categories. The 2 response categories, labeled as **1** and **2** are assumed to be disjoint (i.e. do not overlap). Table 1 describes the outcome of this reliability experiment. It follows from Table 1 that of all N subjects, A are classified by both raters into category 1, while B subjects were classified into category 2 by rater A and in category 1 by rater B. It should be noted that B1 and B2 represent the number of subjects that rater B classified in categories 1 and 2 respectively. A1 and A2 are also defined in a similar way.

Table 1: Distribution of Subjects
By Rater and Response Category

Rater B	**Rater A**		
	1	**2**	**Total**
1	A	B	**B**1=A+B
2	C	D	**B**2=C+D
Total	**A**1=A+C	**A**2=B+D	N

3. SCOTT'S π-STATISTIC

Scott (1955) suggested computing the extent of agreement between raters A and B using the π-statistic *PI*, which is defined as follows:

$$\mathrm{P}I = \frac{p - e(\pi)}{1 - e(\pi)}, \qquad (1)$$

where $p = (A + D)/N$ is the overall agreement propensity and $e(\pi)$ (should be read e of PI) is given by

$$e(\pi) = \left(\frac{(A1 + B1)/2}{N}\right)^2 + \left(\frac{(A2 + B2)/2}{N}\right)^2, \qquad (2)$$

It should be noted that $e(\pi)$ designates the propensity for both raters to agree by chance without having the same assessment of a subject.

The p component of equation (1) only involves subjects both raters have classified in the same category. It could be used as a naïve measure of the extent of agreement. However, there are reasons to believe that raters A and B would classify some subjects into the same category not for the same reasons. Subjects classified in the same category for different reasons correspond to an agreement by chance. Because chance agreement does not measure consistency in the rating, it is not of interest and the p component should be adjusted accordingly. Gwet (2001) discusses extensively about the motivation of the form of equation (1) and explains why statistics of this type provides the desired adjustment.

Unfortunately, identifying the subjects that have led to an agreement by chance is impossible. Therefore, it is necessary to compute the chance of occurrence of an agreement due to luck. This aspect of the problem has led to controversial proposals from researchers and biostatisticians. Scott (1955) recommended equation (2) as a measure of the chance-agreement probability. The first term of equation (2) represents an estimation of the chance that raters A and B independently classify a subject into category 1. The second term on the other hand, estimates the probability for the 2 raters to independently classify a subject into category 2.

4. COHEN'S KAPPA STATISTIC

Cohen (1960) criticized Scott's approach for computing chance-agreement probability because it combines raters A and B classification data. In fact, the first term of equation (2) is obtained by averaging the proportions of subjects classified into category 1 by both raters and raising the average to the power of 2. This approach eliminates any difference that may exist in the rating pattern of both raters.

Cohen (1960) suggested the following "Kappa" statistic for evaluating the extent of agreement between 2 raters:

$$KAPPA = \frac{p - e(\kappa)}{1 - e(\kappa)}, \qquad (3)$$

where $p = (A + D)/N$ and $e(\kappa)$ (read "e of kappa") is defined as follows:

$$e(\kappa) = \left(\frac{A1}{N}\right)\left(\frac{B1}{N}\right) + \left(\frac{A2}{N}\right)\left(\frac{B2}{N}\right). \qquad (4)$$

The first term of equation (4) estimates the chance that both raters independently classify a subject into category 1. The second term on the other hand, estimates the probability of independent classification of a subject into category 2. Unlike Scott's equation (2), the terms in equation (4) are obtained by multiplying individual raters' classification rates.

Despite the differences between the PI and KAPPA statistics in the way chance-agreement probability is estimated, these 2 methods often yield similar results in practice. However, the PI statistic generalizes to the case of multiple raters and multiple-item response categories in a more natural way than the Kappa statistic. Fleiss (1971) proposed a generalization of Scott's PI statistic that is often referred to as KAPPA statistic.

5. LIMITATIONS OF THE KAPPA STATISTIC

Let us consider 2 hypothetical reliability experiments E1 and E2. For each of the experiments, raters A and B have to classify 100 subjects into one of 2 possible response categories, labeled as "1" and "2". Tables 2 and 3 show the outcomes of experiments E1 and E2 respectively.

Table 2: Outcome of Experiment E1

Rater B	**Rater A**		
	"1"	**"2"**	**Total**
"1"	40	9	49
"2"	6	45	51
Total	46	54	100

Table 3: Outcome of Experiment E2

Rater B	**Rater A**		
	"1"	**"2"**	**Total**
"1"	80	10	90
"2"	5	5	10
Total	85	15	100

It should be noted that in both experiments, raters A and B agreed about the classification of 85 subjects. Therefore, the overall agreement propensity p (the p component) is equal to 0.85 (=85/100) for both experiments. One would naturally expect a high inter-rater reliability between the raters in both situations. Unfortunately neither Scott's PI-statistic nor Cohen's statistic provides a consistent extent of agreement in both experiments.

Scott's PI statistic for both experiments is obtained as follows:

- For experiment E1, chance-agreement probability $e(\pi)$ is given by:

  $$e(\pi)=\left(\frac{46+49}{2\times100}\right)^2+\left(\frac{51+54}{2\times100}\right)^2=0.50125.$$

  This leads to a PI statistic of
  PI =(0.85-0.50125)/(1-0.50125) = 0.6993
- For experiment E2, chance-agreement probability is given by:

  $$e(\pi)=\left(\frac{90+85}{2\times100}\right)^2+\left(\frac{10+15}{2\times100}\right)^2=0.78125.$$

  This yields a PI statistic of
  PI = (0.85-0.78125)/(1-0.78125) = 0.3143.

Cohen's KAPPA statistic on the other hand, is obtained as follows:

- For experiment E1, chance-agreement probability $e(\kappa)$ is given by:

  $$e(\kappa)=\left(\frac{49}{100}\right)\times\left(\frac{46}{100}\right)+\left(\frac{51}{100}\right)\times\left(\frac{54}{100}\right)=0.5008.$$

  This leads to a PI statistic of
  KAPPA =(0.85-0.5008)/(1-0.5008) = 0.6995
- For experiment E2, chance-agreement probability is given by:

  $$e(\kappa)=\left(\frac{90}{100}\right)\times\left(\frac{85}{100}\right)+\left(\frac{10}{100}\right)\times\left(\frac{15}{100}\right)=0.78.$$

  This yields a PI statistic of
  KAPPA = (0.85-0.78)/(1-0.78) = 0.318.

The PI and KAPPA statistics surprisingly indicate a low level of agreement between raters A and B following the second experiment. It is in fact difficult to explain why the raters would have a high level of agreement in experiment E1 and a fairly low agreement in experiment E2. This paradox has led several authors to conclude that the Kappa statistic was dramatically affected by the trait prevalence in the population under consideration. Other scientists recommended the testing of marginal homogeneity to determine the adequacy of the KAPPA statistic.

We believe that there are serious conceptual flaws in both statistics (KAPPA and PI) that make them very unreliable. The next section is devoted to the discussion of these flaws.

6. ORIGINS OF INADEQUACY OF KAPPA AND PI STATISTICS

The general form of the Kappa statistic and that of the PI statistic as a function of p and e is appropriate for correcting the agreement propensity for chance agreement. However, it is the expression used to compute the probability of agreement by chance that is inappropriate.

In order to obtain a good equation of chance-agreement probability, it is necessary to define what chance agreement is and to explain the circumstances under which it occurs. Any agreement between 2 raters A and B can be considered as a chance agreement if a rater has performed a random rating (i.e. classified a subject without being guided by its characteristics) and both raters have agreed. If a rating is random, it is possible to demonstrate that agreement can occur with a fixed probability of 0.5. Simulations that we have conducted also tend to confirm this fact. It follows that a reasonable value for chance-agreement probability should not exceed 0.5.

Figures 1 and 2 show a plot of chance-agreement probability for PI and KAPPA statistics, as a function of raters' category-1 marginal classification probabilities. The marginal probabilities are P1A=A1/N for rater A and P1B=B1/N for rater 2. It follows from figure 1 that the chance-agreement probability $e(\pi)$ for the PI-statistic varies from 0.5 to 1. This property seriously contradicts the finding of the previous paragraph, which suggests that, a reasonable value for chance-agreement probability should not exceed 0.5. In fact, all values of $e(\pi)$ exceed 0.5. It is difficult to imagine circumstances where 2 raters would agree by chance with a probability of 1. Moreover, figure 1 also suggests that if A1=0 and B1=0, then raters A and B will agree by chance with probability 1.

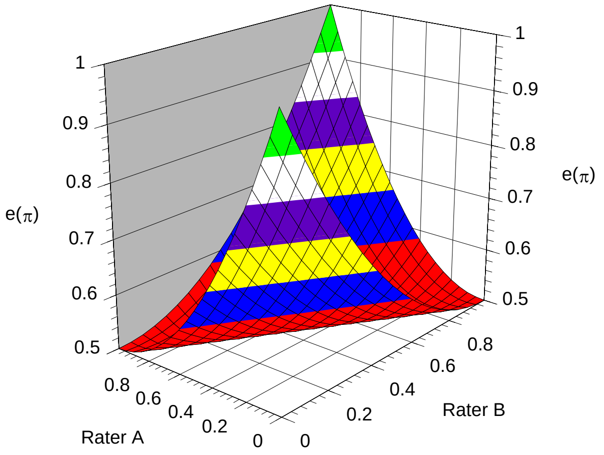

Figure 1: $e(\pi)$ *as a function of raters' classification probabilities*.

Figure 2, shows the graph of $e(\kappa)$ as a function of the marginal classification probabilities P1A and P1B. This figure shows that chance-agreement probability $e(\kappa)$ may take any value between 0 and 1, which violates the condition of being below the upper bound of 0.5. The value of $e(\kappa)$ will generally be smaller than 0.5 if sum of the marginal classification probabilities is reasonably close to 1.

Figure 2 also indicates that the worse Kappa statistics are expected when both marginal classification probabilities P1A and P1B are either very small or very small. These are situations where the curve of $e(\kappa)$ gets closer to its maximum value of 1.

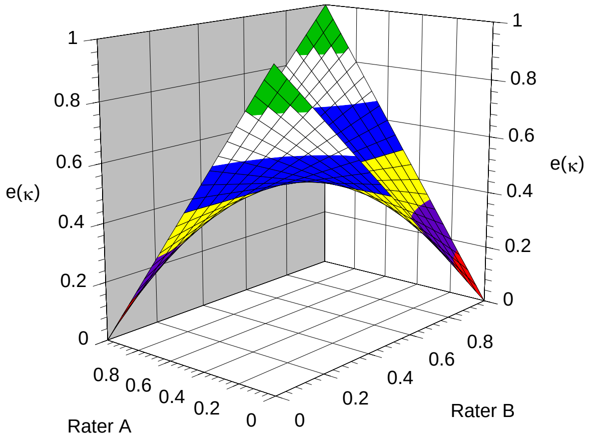

Figure 2: $e(\kappa)$ *as a function of raters' classification probabilities*

7. ALTERNATIVE CHANCE-CORRECTED STATISTIC TO KAPPA

A simple alternative more reliable statistic could be used to estimate the extent of agreement between raters. In the previous section, we considered chance agreement as being a simultaneous occurrence of random rating (by one of the raters) and rater agreement. Therefore, we will compute chance-agreement probability by multi-plying the propensity of random rating and that of agreement under the assumption of random rating.

The propensity of rater agreement under the assumption of random rating is 0.5. That is, if at least one of the 2 raters perform a random classification, they will reach an agreement 50% of the times. This statement can be established mathematically and verified with a simulation. The propensity of random rating on the other hand is defined as the proportion of the maximum classification variance observed in the current reliability experiment. Interested readers are encouraged to read Gwet (2001) for a comprehensive discussion of these concepts.

Let $e(\gamma)$ be the new chance-agreement probability. We have that:

$$e(\gamma) = 2P_1(1-P_1) \qquad (4)$$

where $P_1 = \frac{(A1+B1)/2}{N}$ represents the approximate chance that a rater (A or B) classifies a subject into category 1. The alternative statistic, which is referred to as the AC1-statistic in Gwet (2001) is given by:

$$AC1 = \frac{p-e(\gamma)}{1-e(\gamma)} \qquad (5)$$

where $p = (A+D)/N$ and $e(\gamma)$ given by equation (4).

For the sake of comparability, figure 3 depicts the distribution of $e(\gamma)$ as a function of both raters' classification probabilities.

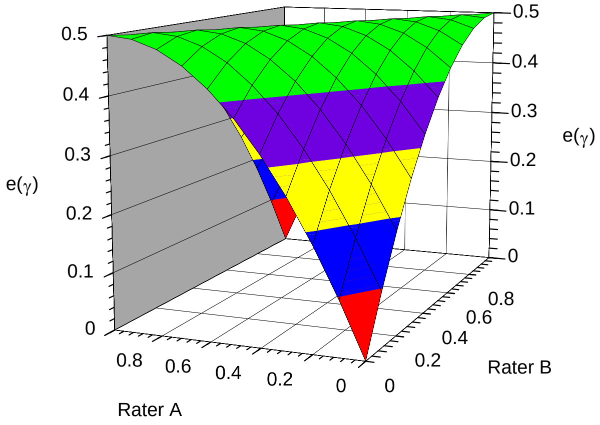

Figure 3: $e(\gamma)$ as a function of raters' classification probabilities

It follows from figure 3 that the probability of chance agreement $e(\gamma)$ always varies between 0 and 0.5. This probability is close to its maximum value of 0.5 if the sum of the marginal probabilities is around 1, and decreases, as this sum gets smaller or bigger than 1.

- For experiment E1 of section 5, chance-agreement probability $e(\gamma)$ is given by: $e(\gamma) = 2\left(\frac{46+49}{2\times 100}\right)\left(1-\frac{46+49}{2\times 100}\right) = 0.49875.$ It follows that the AC1 statistic is given by:
  AC1 = (0.85-0.49875)/(1-0.49875)
  = 0.7008
  This inter-rater reliability is very similar to those obtained with the PI and KAPPA statistics.
- For experiment E2 on the other hand, chance-agreement probability $e(\gamma)$ is obtained as follows:
  $$e(\gamma) = 2\left(\frac{90+85}{2\times 100}\right)\left(1-\frac{90+85}{2\times 100}\right) = 0.21875.$$
  The AC1 statistic for this experiment is given by:
  AC1 = (0.85-0.21875)/(1-0.21875)
  = 0.808.
  For this experiment, the inter-rater reliability was estimated at 0.3143 and 0.318 with the PI and KAPPA statistics respectively. It appears clearly that the AC1 statistic provides an estimate that is more consistent with the outcome of experiment E2 as described in Table 3. An inter-rater reliability as low as 0.32 obtain from Table 3 is difficult to justify.

8. CONCLUSION

The objective of this article was two-fold:

(1) To prove that the widely-used KAPPA and PI statistics can be misleading in many cases, especially when the sum of the marginal probabilities is very different from 1.

(2) To introduce an alternative more robust chance-corrected statistic that consistently yields reliable results.

We have established that the unpredictable behavior of the PI and KAPPA statistics is due to a wrong method of computing chance-agreement probability. This has unfortunately led some researchers to question the very merit of chance-corrected statistics for estimating inter-rater reliability. The AC1 statistic provides a more reliable approach. Methods for testing the AC1 statistic for statistical significance are given in Gwet (2001), where a detailed motivation of the AC1 statistic and a generalization to the case of multiple raters are presented.

9. REFERENCES

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. (2001). *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
STATAXIS Consulting
15914B Shady Grove Road, PMB 145
Gaithersburg, MD 20877.
U.S.A.
Kilem62@yahoo.com