Kappa and Beyond: Is There Agreement?

Joseph R. Dettori, PhD¹ and Daniel C. Norvell, PhD¹

Introduction

It is often of interest to assess agreement between 2 or more observers when the observation of interest is categorical. Observations of clinical interest can include diagnoses, assessment of risk factors (exposures), and outcomes. The statistical tool most widely used to determine agreement is the kappa statistic. The kappa statistic is a chance-corrected measure; that is, it seeks to assess agreement beyond that which occurs by chance. However, it is well known that kappa has some limitations that can lead to paradoxical results (low kappa even in the presence of strong observer agreement) in certain circumstances. In this article, we will review kappa and its limitations, and introduce an alternative measure of agreement, the Agreement Coefficient 1 (AC₁) given by Gwet.

The Kappa Statistic

Cohen in 1960 proposed the kappa statistic in the context of 2 observers. It was later extended by Fleiss to include multiple observers. For illustration purposes, we will look at the simpler case of 2 observers, acknowledging that the principles are the same for multiple observers. Let us imagine that an investigator wants to develop a spine classification system based in part on the presence or absence of a finding on magnetic resonance imaging (MRI). The investigator gives 200 MRIs to 2 spine surgeons to assess independently whether the finding is present or absent in each image. The results are provided in Table 1.

The kappa statistic is given by the formula

\[ \kappa = \frac{P_o - P_e}{1 - P_e} \]

where \( P_o \) = observed agreement, \( (a + d)/N \), and \( P_e \) = agreement expected by chance, \( (\frac{g_1 \times f_1}{N}) + (\frac{g_2 \times f_2}{N}) \).

In our example,

\[ P_o = \frac{(130 + 5)}{200} = 0.675 \]
\[ P_e = \frac{((186 \times 139) + (14 \times 61))}{200^2} = 0.668 \]
\[ \kappa = \frac{0.675 - 0.668}{1 - 0.668} = 0.022 \]

Kappa values range from -1 to 1, though it usually falls between 0 and 1. One represents perfect agreement, indicating that the raters agree in their classification of every case. Zero indicates agreement no better than that expected by chance. A negative kappa would indicate agreement worse than that expected by chance.

A common scale of interpretation for the kappa statistic is given by Altman (Table 2). Using the scale shown in Table 2, the ability of our 2 investigators to agree on the presence or absence of a specific MRI finding is “poor.”

The Limitations of Kappa

While the kappa statistic above is quite low, one may notice that the absolute percentage of observer agreement (\( P_o \)) is quite

---

Table 1. A 2 x 2 Table of Results in a Hypothetical Example Comparing 2 Different Assessors.

| Present | Absent | Total |
|---------|--------|-------|
| **Assessor 1** | **Assessor 2** |
| (a) 130 | (c) 9 |
| (b) 56 | (d) 5 |
| **Total** | **Total** |
| (g1) 186 | (g2) 14 |
| (f1) 139 | (f2) 61 |
| (N) 200 | (N) 200 |

Table 2. A Commonly Used Scale of Interpretation for Kappa Statistic.

| Kappa       | Agreement |
|-------------|-----------|
| ≤0.20       | Poor      |
| 0.21-0.40   | Fair      |
| 0.41-0.60   | Moderate  |
| 0.61-0.80   | Good      |
| 0.81-1.00   | Very good |

¹ Spectrum Research, Inc, Steilacoom, WA, USA

Corresponding Author:
Joseph R. Dettori, Spectrum Research, Inc, PO Box 88998, Steilacoom, WA 98388, USA.
Email: joe@specri.com

Creative Commons Non Commercial No Derivs CC BY-NC-ND: This article is distributed under the terms of the Creative Commons Attribution-Non Commercial-NoDerivs 4.0 License (https://creativecommons.org/licenses/by-nc-nd/4.0/) which permits non-commercial use, reproduction and distribution of the work as published without adaptation or alteration, without further permission provided the original work is distributed as specified on the SAGE and Open Access pages (https://us.sagepub.com/en-us/nam/open-access-at-sage).
The magnitude of chance agreement (lies in the distribution of the marginal table totals on which the time yet have such a low kappa? The answer to this question ¼ comes from the agreement expected by chance as follows:

\[ q = \frac{(g_1 + f_1)}{2N} \]

In our example, \( P_o \) remains the same, but \( P_e \) takes on a different value with the following results:

\[ P_o = \frac{(130 + 5)}{200} = 0.675 \]
\[ q = \frac{(186 + 139)}{2 \times 200} = 0.813 \]
\[ P_e = 2(0.813) \times (1 - 0.813) = 0.305 \]
\[ AC_1 = \frac{(0.675 - 0.305)}{(1 - 0.305)} = 0.532 \]

How shall we interpret this coefficient? Remember from Table 2 that many use a common benchmark scale to determine if the agreement is poor, fair, moderate, good, or very good. The advantage is that this scale is straightforward. However, this simple method can lead to misleading conclusions for the following reasons:

- The calculated kappa is specific to the pool of subjects used in the study and will change if different subjects are used.
- The magnitude of the kappa coefficient is dependent on several factors such as sample size, the number of categories, and the distribution of subjects among the categories. For example, a kappa value of 0.54 based on 200 subjects suggests a much stronger message about the extent of agreement among raters, than a kappa value of 0.6 based on 10 subjects only.

A more standardized method that overcomes many of the problems cited above was proposed by Gwet. Using the agreement coefficient and its standard error, one can calculate the probability that the agreement coefficient would fall into each category in Table 2. Starting with the highest agreement coefficient range, 0.81 to 1.0 (very good), and moving to the poorest range (≤0.2), one calculates the cumulative probability that the coefficient falls into that category. When the cumulative probability crosses a certain threshold (say 95%) that is the most likely range to which the estimate belongs. The results from our example show that kappa’s cumulative probability does not reach 95% until the bottom range (poor), where the AC1 crosses in the 0.41 to 0.60 range (moderate; (Table 3). Fortunately, Gwet’s agreement coefficient and his method for benchmarking are included in several statistical packages.

### An Alternative Statistic of Agreement

Given that kappa is affected by the skewed distributions of categories (the prevalence problem) and by the degree to which observers disagree (the bias problem), Gwet in 2008 proposed an alternative agreement statistic by adjusting for chance agreement in a different way. He defined and new agreement coefficient (AC1) between 2 (or multiple) observers as the conditional probability that 2 randomly selected observers will agree, given that no agreement will occur by chance. In this way, the AC1 resists the so-called “paradox” of kappa. The AC1 statistic has the same formula as the kappa except it calculates the agreement expected by chance as follows:

\[ P_e = \text{agreement expected by chance}, 2q \times (1 - q), \text{ where} \]
\[ q = \frac{(g_1 + f_1)}{2N} \]

In our example, \( P_o \) remains the same, but \( P_e \) takes on a different value with the following results:

\[ P_o = \frac{(130 + 5)}{200} = 0.675 \]
\[ q = \frac{(186 + 139)}{2 \times 200} = 0.813 \]
\[ P_e = 2(0.813) \times (1 - 0.813) = 0.305 \]
\[ AC_1 = \frac{(0.675 - 0.305)}{(1 - 0.305)} = 0.532 \]

### Conclusions

- Kappa remains the most frequently used statistic assessing agreement between 2 or more observers when the observation of interest is categorical.
- The kappa statistic is a chance-corrected measure. However, there are some limitations to the kappa that relate to the distribution of the marginal table totals on which the chance correction depends.
• An alternative agreement statistic is Gwet’s AC1 that seeks to minimize the kappa limitations.
• A standardized method of benchmarking includes calculating the cumulative probability of an agreement coefficient falling within a benchmark range, providing a more standardized way of interpreting the agreement statistic.

Declaration of Conflicting Interests
The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding
The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Spectrum Research, Inc, received funding in exchange for assistance in providing methodological and statistical assistance.

References
1. Cicchetti DV, Feinstein AR. High agreement but low kappa: II. Resolving the paradoxes. J Clin Epidemiol. 1990;43:551-558.
2. Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes. J Clin Epidemiol. 1990;43:543-549.
3. Gwet KL. Computing inter-rater reliability and its variance in the presence of high agreement. Br J Math Stat Psychol. 2008;61(pt 1):29-48.
4. Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull. 1968;70:213-220.
5. Fleiss JL. Measuring nominal scale agreement among many raters. Psychol Bull. 1971;76:378-382.
6. Altman DG. Practical Statistics for Medical Research. London, England: Chapman & Hall/CRC; 1991.
7. Gwet KL. K.Gwet’s Inter-Rater Reliability Blog. Benchmarking agreement coefficients https://inter-rater-reliability.blogspot.com/2014/. Published December 12, 2014. Accessed January 3, 2020.