CHAPTER 4

AC_1 and α Coefficients

OBJECTIVE

This chapter presents two alternatives to Kappa named the AC_1 and Aickin's α (not to be confounded with Krippendorff's α of the previous chapter) proposed by Gwet (2008*a*) and Aickin (1990) respectively, the focus being on AC_1. Both coefficients were developed to overcome Kappa's paradox problem discussed in chapter 2. The probabilistic models underlying these 2 coefficients are discussed as well. Also introduced is Gwet's AC_2, the extension of AC_1 to ordinal, interval and ratio ratings

CONTENTS

"There is no true value of any characteristic, state, or condition that is defined in terms of measurement or observation. Change of procedure for measurement (change in operational definition) or observation produces a new number $\cdots$. There is no such thing as a fact concerning an empirical observation."

- Edwards Deming (1900-1993) -

4.1 Overview

We devote this chapter primarily to the AC_1 statistic proposed by Gwet (2008*a*) as a paradox-resistant alternative to the unstable Kappa coefficient. Will also be discussed is the alpha (α) coefficient of Aickin (1990), an inter-reliability statistic based on a clear-cut definition of the notion of "extent of agreement among raters." Both coefficients differ from Kappa, mainly in the way the percent chance agreement is calculated. As a matter of fact, the poor statistical properties of Kappa stem from an inadequate approach used for computing the percent chance agreement. The Kappa and Pi coefficients rely on a chance-agreement probability expression that is valid only under the improbable assumption that all ratings are known to be independent even before the experiment had been carried out. To justify the 2 expressions used to evaluate the chance-agreement probabilities of Kappa and Pi, the reasoning was that if the processes by which 2 raters classify a subject are statistically independent, then the probability that they agree is the product of the individual probabilities of classification into the category of agreement. However, raters often rate the same subjects, and are therefore expected to produce ratings that are dependent with possibly a few exceptions.

Throughout this chapter, we will consider that independence occurs when a nondeterministic[1] rating is assigned to a subject that is hard to rate. Nondeterministic ratings may be expected on a small fraction of the subject population only, and certainly not on the whole population. The AC_1 of Gwet(2008*a*), and the alpha of Aickin (1990) are based upon the more realistic assumption that only a portion of the observed ratings will potentially lead to agreement by chance. The difficulty to overcome will be to estimate the percent of subjects that are associated with a nondeterministic rating.

When we started the work on improving Kappa, we were not aware of Aickin's theory until after the publication of the ideas to be discussed here in Gwet (2008*a*). After studying Aickin's work, we discovered that our proposed framework was more general, and that the conceptual definition of the extent of agreement among raters

[1]The process of rating a subject is considered *nondeterministic* if it has no apparent connection with the subject's characteristics.

we proposed was also different from Aickin's. Aickin's alpha coefficient for 2 raters represents the portion of the entire population of subjects that both raters will classify identically for cause, as opposed to classifying them identically by chance. To see what Gwet's AC_1 for 2 raters conceptually represents, imagine that all subjects to be classified into identical categories by pure chance are first identified, then removed from the population of subjects. This operation creates a new trimmed population where agreement by chance would be impossible. The AC_1 coefficient is the relative number of subjects in the trimmed subject population upon which the raters agreed. AC_1 and alpha coefficients both represent a probability of agreement for cause, which are calculated with respect to 2 different baseline subject populations. Although limited to the case of two raters only, we have found Aickin's proposal useful and decided to include it in the discussions.

Among Kappa's strengths is a genuine attempt to correct the percent agreement for chance agreement, and the simplicity with which this was done. Among its limitations are the paradoxes described by Feinstein and Cicchetti (1990), where Kappa would yield a low value when the raters show high agreement. In this chapter we propose the AC_1 coefficient, which is similar to Kappa in its formulation and its simplicity, in addition to being paradox-resistant. The alpha coefficient is also close to Kappa in its form. But unlike Kappa and AC_1, the alpha coefficient is computation-intensive with its iterative procedure. AC_1 and alpha both share the same feature of being paradox-resistant.

4.2 Gwet's AC_1 and Aickin's α for 2 Raters

This section describes the procedures for computing the AC_1 and α coefficients in the case of 2 raters classifying a sample of n raters into one of q possible categories. The calculation of these coefficients will also be illustrated in a numerical example.

4.2.1 *The AC_1 Statistic*

Let us consider a two-rater reliability experiment based on a q-level nominal measurement scale. As previously indicated, rating data resulting from such an experiment could be conveniently organized in a contingency table such as Table 2.7 in chapter 2. The AC_1 coefficient, denoted by $\widehat{\gamma}_1$ is defined as follows :

$$\widehat{\gamma}_1 = \frac{p_a - p_e}{1 - p_e}, \text{ with } p_a = \frac{1}{1 - p_{\text{M}}} \sum_{k=1}^{q} p_{kk}, \; p_e = \frac{1}{q-1} \sum_{k=1}^{q} \pi_k (1 - \pi_k), \tag{4.1}$$

where p_{M} is the relative number of subjects rated by a single rater (i.e. 1 rating is missing)[2], and $\pi_k = (p_{k+} + p_{+k})/2$. Note that p_{k+} and p_{+k} represent the relative number of subjects assigned to category k by raters A and B respectively. The symbol p_{kk} is the relative number of subjects classified into category k by both raters. While π_k represents the probability of a randomly-selected rater to classify a randomly-selected subject into category k, the chance-agreement probability p_e is a product of the following 2 quantities:

- The probability that 2 raters agree given that the subject being rated was assigned a nondeterministic score. This conditional[3] probability is $1/q$.
- The propensity for a rater to assign a nondeterministic score, which is estimated by the ratio: $\sum_{k=1}^{q} \pi_k(1-\pi_k)/(1-1/q)$.

Section 4.4 contains a more detailed discussion of the theory behind this statistic. Gwet (2008*a*) also provides examples and theoretical results related to the AC_1 statistic.

4.2.2 *Aickin's α-Statistic*

The alpha statistic $\widehat{\alpha}$ of Aickin (1990) is defined as follows:

$$\widehat{\alpha} = \frac{p_a - p_e}{1 - p_e}, \text{ where } p_e = \sum_{k=1}^{q} p_{k|\text{H}}^{(\text{A})} \cdot p_{k|\text{H}}^{(\text{B})}, \tag{4.2}$$

and $p_{k|\text{H}}^{(\text{A})}$ represents the probability for rater A to classify into category k, a subject known to be hard to classify. The final classification of this particular group of hard-to-classify subjects involves guesswork and will be random. The overall percent agreement p_a is the same as that of equation 4.1. The main difference between Kappa and alpha lies in the way the percent chance agreement is calculated. While Kappa's percent chance agreement includes all ratings, Aickin's only uses ratings associated with the hard-to-classify subjects. Aickin's theory from which the alpha coefficient is derived, is discussed in section 4.3.

Because the group of hard-to-classify subjects is not identifiable, there is no simple expression for obtaining the probabilities $p_{k|\text{H}}^{(\text{A})}$ and $p_{k|\text{H}}^{(\text{B})}$. To solve this problem, Aickin (1990) proposed an iterative algorithm based on the following system of 3 equations:

[2] All subjects not rated by either rater are excluded from the analysis.

[3] The condition here being the nondeterministic nature of the ratings, which will lead any resulting agreement to be considered chance agreement.

$$\hat{\alpha}^{(t+1)} = \frac{p_a - p_e^{(t)}}{1 - p_e^{(t)}}, \ \textit{where } p_e^{(t)} = \sum_{k=1}^{q} p_{k|\mathrm{H}}^{\mathrm{A}(t)} \cdot p_{k|\mathrm{H}}^{\mathrm{B}(t)}, \tag{4.3}$$

$$p_{k|\mathrm{H}}^{\mathrm{A}(t+1)} = \frac{p_{k+}}{\left(1 - \hat{\alpha}^{(t)}\right) + \hat{\alpha}^{(t)} p_{k|\mathrm{H}}^{\mathrm{B}(t)} \Big/ p_e^{(t)}}, \ \textit{for } k = 1, \cdots, q, \tag{4.4}$$

$$p_{k|\mathrm{H}}^{\mathrm{B}(t+1)} = \frac{p_{+k}}{\left(1 - \hat{\alpha}^{(t)}\right) + \hat{\alpha}^{(t)} p_{k|\mathrm{H}}^{\mathrm{A}(t)} \Big/ p_e^{(t)}}, \ \textit{for } k = 1, \cdots, q. \tag{4.5}$$

This iterative process is initiated with the marginal probabilities p_{k+} and p_{+k} as starting values for the varying probabilities $p_{k|\mathrm{H}}^{\mathrm{A}(t)}$ and $p_{k|\mathrm{H}}^{\mathrm{B}(t)}$. That is, $p_{k|\mathrm{H}}^{\mathrm{A}(0)} = p_{k+}$, and $p_{k|\mathrm{H}}^{\mathrm{B}(0)} = p_{+k}$. Therefore, the initial alpha value $\hat{\alpha}^{(0)}$ when $t = 0$ will be identical to the classical Kappa statistic. The next alpha value $\hat{\alpha}^{(1)}$ when $t = 1$ is calculated from $\hat{\alpha}^{(0)}$ and the other probability values according to the above equations. The iterative process will stop when 2 consecutive Alpha values $\hat{\alpha}^{(t+1)}$ and $\hat{\alpha}^{(t)}$ does not exceed a predetermined small value such as 0.001, which depends on the precision you like to achieve.

4.2.3 *Example*

This section presents a practical example to illustrate the calculation of the AC_1 and α agreement coefficients. To compute the α coefficient, Aickin recommends to add a pseudo-count[4] of 1 to the total count of subjects, and to distribute it uniformly among all cells to avoid convergence problems with the iterative algorithm.

Example 4.1

To illustrate the calculation of AC_1 and alpha coefficients, let us consider the reliability data of Table 4.1. This data represents the distribution of human subjects suffering from back pain, by pain type, and observing clinician.

Table 4.1:
Ratings of Spinal Pain by Clinicians 1 and 2, and Pain Type

Clinician 1	Clinician 2		
	Derangement Syndrome	Dysfunctional Syndrome	Postural Syndrome
Derangement Syndrome	55	10	2
Dysfunctional Syndrome	6	4	10
Postural Syndrome	2	5	6

[4] A "pseudo-count" is an integer value primarily used for changing artificially a cell count value from being 0 to being negligible. Zero-count cells are known to be problematic to probability-based computing systems, but cannot be eliminated unless they represent events known to be impossible.

Cohen's Kappa for this data is given by $\widehat{\kappa}_{\text{C}} = (0.65-0.4835)/(1-0.4835) = 0.3224$. The AC_1 coefficient on the other hand is $\widehat{\gamma}_1 = (0.65-0.257725)/(1-0.257725) = 0.5285$. As for Aickin's Alpha, after 10 iterations we obtained $\hat{\alpha} = (0.65-0.4121)/(1-0.4121) = 0.4047$, and the final "marginal" probabilities related to hard-to-classify subjects are given by $(p^{(\text{A})}_{1|\text{H}}, p^{(\text{A})}_{2|\text{H}}, p^{(\text{A})}_{3|\text{H}}) = (0.5993437,\ 0.2442839,\ 0.1563717)$ for clinician A, and by $\left(p^{(\text{B})}_{1|\text{H}}, p^{(\text{B})}_{2|\text{H}}, p^{(\text{B})}_{3|\text{H}}\right) = (0.5321665,\ 0.2274873,\ 0.2403553)$ for clinician B.

The Kappa, alpha, and AC_1 statistics in example 4.1 are respectively given by 0.322, 0.405, and 0.529. Kappa represents less than half the magnitude of the overall agreement probability $p_a = 0.65$. This drastic reduction in the magnitude of p_a results from Kappa's unduly high chance-agreement correction. AC_1 on the other hand accounts for more than 80% of that value. As will be seen in the next few sections, alpha does not measure the same concept as Kappa and AC_1 and a direct comparison would be inappropriate.

While AC_1 and Kappa represent agreement probabilities based on the pool of subjects from which the Hard-to-classify ones have been removed, alpha on the other hand represents the probability of "for-cause" agreement[5] based on all subjects. Because the reference population for evaluating alpha is bigger, $\hat{\alpha}$ will generally be lower than AC_1 unless the group of subjects does not have those special subjects who may lead to an agreement by chance. In practice alpha will often be greater than Kappa, and this is primarily due to the poor performance of Kappa in many situations.

The next two sections 4.3 and 4.4 deal with the theoretical foundations of the alpha and AC_1 statistics and require some limited abstract thinking. Our primary objective in these 2 sections is to answer the following question : "*If we knew everything about all subjects, and raters of interest (including the ratings that the raters would assign to each rater, and the raters' skill level), how would we evaluate inter-rater reliability ?*" This hypothetical situation will lead to the creation of a theoretical framework. In a real experiment based on a sample of subjects, some aspects of this problem, which are taken for granted within the theoretical framework will be unknown. We will then need to resort to some estimation procedures to compensate for these gaps in our knowledge. The result will be a statistical procedure subject to sampling errors, which will be studied with the techniques of inferential statistics discussed in Chapter 5.

Although sections 4.3 and 4.4 explain the motivation behind the formulation of AC_1, and that of alpha, they are not essential for using equations 4.1 and 4.2 in

[5] A "for-cause" agreement is an agreement situation where both raters classified a subject into the same category for a reason.