CHAPTER 4

Constructing Agreement Coefficients: AC_1 and Aickin's α

OBJECTIVE

This chapter presents a detailed discussion of two paradox-resistant alternative agreement coefficients 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. These two agreement coefficients will be constructed step by step, from the definition of the theoretical construct to the formulation of the coefficient. All intermediary steps, which include the underlying statistical model, and the subject and rater population parameters will be spelled out. This chapter focuses particularly on the AC_1 coefficient, and aims at providing a detailed account of its real meaning, its advantages, and possible limitations. Also discussed 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

In this chapter, I discuss two particular agreement coefficients: (1) the AC_1 statistic proposed by Gwet (2008*a*) as a paradox-resistant alternative to the unstable Kappa coefficient, and (2) the alpha (α) coefficient of Aickin[1] (1990), an inter-reliability statistic based on a clear-cut definition of the notion of "extent of agreement among raters." I present the reader with a clear view of a step-by-step construction of an agreement coefficient, and will conduct an elaborate discussion of the underlying assumptions. Both coefficients differ from Kappa mainly in the way the percent chance agreement is calculated. As a matter of fact, the notion of chance agreement is pivotal in the study of chance-corrected agreement coefficients. Understanding it well is essential for developing effective agreement coefficients. The poor statistical properties of Kappa for example stem precisely from the inadequate approach used to evaluate the percent chance agreement.

Several authors have justified the Kappa coefficient on the ground that it represents the difference between the observed percent agreement (p_a) and the percent chance agreement[2] (p_e), which is normalized by its maximum value ($1 - p_e$) so that the coefficient is confined within the $(0, 1)$ interval. The problem is that this whole operation describes something that may not even be remotely close to what raters actually do. My views on this are more in line with Grove et al. (1981) who while talking about what diagnosticians in the medical field actually do said this: *"They assign the easy cases or textbook cases, to diagnoses with little or not error; they may guess or diagnose randomly on the others. If one knew which cases were textbook cases, one could them separately; but that is a difficult matter."* I strongly believe that the distinction between textbook and non-textbook cases is the crux of the matter. Confronting this issue head-on is as important and difficult as it is inevitable, and how it is approached might decide how good or bad the agrement coefficient will turn out to be.

Grove et al. (1981) describes Kappa's percent chance agreement in the following

[1] Not to be confounded with Krippendorff's alpha, which is an entirely different coefficient discussed in the previous chapters.

[2] Chance agreement here stands for agreement when two raters assign ratings to subjects randomly.

terms: "When in doubt on a nontextbook case, each rater mentally flips a biased coin, with the probability of getting heads (giving the diagnosis) equal to his own base rates ..." This characterization of Kappa's percent chance agreement is likely too generous because Kappa's percent chance agreement does not behave near as well. The problem stems from the first 3 words "When in doubt" of this quote. In fact, there nothing considered to be an integral part of kappa, which suggests that its expression for chance-agreement probability applies only when the raters are in doubt. Kappa expression does not incorporate an estimate of the nontextbook or uncertain cases.

The Kappa and Pi coefficients rely on a percent chance agreement or 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 two expressions used to evaluate the chance-agreement probabilities of Kappa and Pi, the reasoning was that if the processes by which two 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 when they are in doubt.

Throughout this chapter, I consider that independence occurs when a nondeterministic[3] rating (generally associated with hard or nontextbook cases) is assigned to a subject that is hard to rate. Nondeterministic ratings may be expected on a small fraction of subjects only, and certainly not on the whole subject sample or 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 I started working on an alternative to the Kappa coefficient, I was unaware of Aickin's work. I learned about it only after the publication in Gwet (2008*a*), of the ideas to be discussed here. I then discovered that the framework I proposed was made more general by allowing the group of textbook subjects to be specific to each rater instead of being unique for all raters as Aickin assumed. Moreover, my conceptual definition of the extent of agreement among raters differ from Aickin's. That is both coefficients do not quantify the same concept. Aickin's alpha coefficient for two raters represents the portion of the entire population of subjects that both raters are expected to classify identically for cause, as opposed to classifying them identically by chance. To see what Gwet's AC_1 for two raters conceptually represents,

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

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 of subjects 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 are expected to agree. AC_1 and alpha coefficients both represent a probability of agreement for cause, which are calculated with respect to two different reference subject populations. Although it is limited to two raters only, I have found Aickin's proposal useful and decided to include it into 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 I propose the AC_1 coefficient, which has some similarities with 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 two Raters

This section describes the procedures for computing the AC_1 and α coefficients in the case of two raters classifying a sample of n subjects 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[4] by $\widehat{\gamma}_1$ is defined as follows:

$$\widehat{\gamma}_1 = \frac{p_a - p_e}{1 - p_e},\ \textit{with}\ p_a = \sum_{k=1}^{q} p_{kk},\ p_e = \frac{1}{q-1}\sum_{k=1}^{q} \pi_k(1-\pi_k), \tag{4.2.1}$$

[4] I use $\widehat{\gamma}_1$ (read "gamma hat one") to designate the value of AC_1 estimated from observed ratings taken on a sample of subjects. Its estimand γ_1 is the AC_1 value based on the entire subject population. Later in this chapter, I will use the symbol $\widehat{\gamma}_2$ to designate AC_2, which the weighted version of AC_1.

Table 4.12: Weighted Agreement Coefficients for Table 4.9 Rating Data

Coefficient	Unweighted	Weighted
Conger's Kappa	0.4762	0.7435
Gwet's AC_1	0.5021	0.8224
Fleiss' Kappa	0.4651	0.7305
Krippendorff's α	0.4817	0.7468
Brenann-Prediger	0.4933	0.7980
Percent Agreement	0.6200	0.9439

4.7 Concluding Remarks

The primary objective of this chapter was to present a theoretical framework for investigating the notion of agreement among raters, and to introduce the AC_1 statistic as a robust alternative agreement coefficient to Kappa, Pi, or Krippendorff's alpha. I wanted to have an in-depth discussion about the relationship between the computational procedures and the concept of agreement for the purpose of justifying the different approaches. I wanted the reader to see why some computational procedures are formulated the way they are, and to see what their limitations may be. The whole framework developed in this chapter is based on the notion of E-subjects or easy subjects also known as textbook subjects, and on the notion of H-subjects or hard subjects that are expected to be rated randomly and susceptible to produce agreement among raters by pure chance. Because of the difficulty to tease apart the subpopulations of E- and H-subjects, it is often necessary to make assumptions in order to obtain a definitive formulation of the computational procedures.

All agreement coefficients discussed in the past few chapters were developed around the percent agreement p_a, which is corrected for chance agreement using various strategies. Agreement by pure chance is perceived by most researchers as false agreement that if left untreated may artificially increase the estimated extent of agreement among raters. Therefore, correcting the percent agreement for chance agreement aims at dampening down the adverse effect of spurious agreements on the agreement coefficients, and the false sense of uniformity in the ratings they convey. An agreement by pure chance does not reflect any leveling in raters' knowledge and skills.

While adjusting for chance agreement is necessary, not all adjustment methods are expected to bring the percent agreement closer to the "true" extent of agreement among raters. Kappa, and Pi are known to behave as well as other alternative

coefficients only when the percent agreement is around 0.5. The BP coefficient's performance appears to be superior to that of Kappa and Pi. However, the fixed percent chance agreement of 0.5 used by BP sometimes indicates a propensity for chance agreement that exceeds what would be expected from the data. This artificially reduces the magnitude of the inter-rater reliability coefficient.

Improving inter-rater reliability coefficients requires one to define a construct and to formulate the operational definition that shapes it, possibly through statistical modeling. Aickin (1990) proposed the approach discussed in section 4.3. His approach led to an agreement coefficient with good statistical properties, and is based on the notion of "Hard-to-score Subjects" who are assigned nondeterministic ratings. Among the disadvantages of Aickin's alpha coefficient are its time-consuming iterative computation procedure, and its magnitude that cannot reach the maximum value of 1. Gwet's approach, which led to the AC_1 coefficient is discussed in section 4.4. It also uses the notion of "Hard-to-score Subjects" that produce agreement by chance. While Aickin's alpha represents the relative number of "Easy-to-score Subjects" with respect to the total number of subjects, Gwet's AC_1 represents the relative number of "Easy-to-score Subjects" with respect to the group of subjects left after removing H-subjects. That is, if the raters agree on all H-subjects then the AC_1 coefficient will be 1, and Aickin's alpha will still be the proportion of E-subjects in the population. In fact alpha takes the maximum value of 1 only if there is no H-subject in the subject population. For Aickin (1990) the very existence of H-subjects makes it impossible to obtain a perfect agreement even if there is no observed disagreement. For Gwet (2008), only an observed disagreement (on an H-subject) would make it impossible to obtain the maximum agreement coefficient of 1.

When special types of disagreements represent a certain level of agreement (or partial agreement), the AC_2 coefficient introduced in section 4.6, provides a more accurate assessment of the inter-rater reliability. This is achieved by assigning a weight to each pair of scores, downweighting the pairs that represent little agreement while upweighting those representing substantial agreement. Although we have only considered a few types of weights in this chapter, practitioners could consider different weights to serve different purposes, provided the weights used in the analysis are defined prior to the reliability experiment.