CHAPTER 5

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 (2008a) 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 · · · . There is no such thing as a fact concerning an empirical observation."*
>
> - Edwards Deming (1900-1993) -

5.1 Overview

In this chapter, I discuss two particular agreement coefficients: (1) the AC_1 statistic proposed by Gwet (2008a) as a paradox-resistant alternative to the unstable Kappa coefficient, and (2) the alpha (α) coefficient of Aickin (1990)[1], 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 (2008a), 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 (2008a), 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, imagine that all subjects to be classified into identical categories by pure chance

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

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.

5.2 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 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.

5.2.1 *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 3.5 in chapter 3. The AC_1 coefficient denoted[4] by $\widehat{\kappa}_{1G}$ is defined as follows:

$$\widehat{\kappa}_{1G} = \frac{p_a - p_e}{1 - p_e}, \text{ with } p_a = \sum_{k=1}^{q} p_{kk}, \; p_e = \frac{1}{q-1}\sum_{k=1}^{q} \pi_k(1-\pi_k), \tag{5.2.1}$$

where $\pi_k = (p_{k+} + p_{+k})/2$. 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}

[4] I use $\widehat{\kappa}_{1G}$ to designate the value of AC_1 estimated from observed ratings taken on a sample of subjects. Its estimand κ_{1G} is the AC_1 value based on the entire subject population. Later in this chapter, I will use the symbol $\widehat{\kappa}_{2G}$ to designate AC_2, which is the weighted version of AC_1.

is the relative number of subjects classified into category k by both raters. While π_k represents the probability for 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 two quantities:

- The probability that two raters agree given that the subject being rated is nontextbook and was therefore assigned a nondeterministic rating. This conditional[5] probability is $1/q$ since nondeterministic ratings are considered random with equal chance for all q categories.

- The propensity for a rater to assign a nondeterministic rating, which is estimated by the ratio: $\sum_{k=1}^{q} \pi_k(1-\pi_k)/(1-1/q)$. More will be said about this expression later in this chapter. What is important to retain from this expression is that a distribution of subjects that is skewed towards a few categories will lower the nondeterministic rating propensity.

Section 5.4 contains a more detailed discussion of the theory behind this statistic. Gwet (2008a) also provides examples and theoretical results related to the AC_1 statistic.

5.2.2 *Aickin's α-Statistic*

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

$$\widehat{\alpha}_A = \frac{p_a - p_e}{1 - p_e}, \text{ where } p_e = \sum_{k=1}^{q} p_{k|H}^{(A)} \cdot p_{k|H}^{(B)}, \tag{5.2.2}$$

and $p_{k|H}^{(A)}$ represents the probability for rater A to classify into category k, a subject known to be hard to classify (i.e. a nontextbook subject). The final classification of this particular group of hard-to-classify subjects involves guesswork and will be random. The percent agreement p_a is the same as that of equation 5.2.1. 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 hard-to-classify subjects. Aickin's theory from which the alpha coefficient is derived, is discussed in section 5.3.

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

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

$$\hat{\alpha}^{(t+1)} = \frac{p_a - p_e^{(t)}}{1 - p_e^{(t)}}, \text{ where } p_e^{(t)} = \sum_{k=1}^{q} p_{k|\mathrm{H}}^{\mathrm{A}(t)} \cdot p_{k|\mathrm{H}}^{\mathrm{B}(t)}, \tag{5.2.3}$$

$$p_{k|\mathrm{H}}^{\mathrm{A}(t+1)} = \frac{p_{k+}}{\left(1 - \hat{\alpha}^{(t)}\right) + \hat{\alpha}^{(t)} p_{k|\mathrm{H}}^{\mathrm{B}(t)} \Big/ p_e^{(t)}}, \text{ for } k = 1, \cdots, q, \tag{5.2.4}$$

$$p_{k|\mathrm{H}}^{\mathrm{B}(t+1)} = \frac{p_{+k}}{\left(1 - \hat{\alpha}^{(t)}\right) + \hat{\alpha}^{(t)} p_{k|\mathrm{H}}^{\mathrm{A}(t)} \Big/ p_e^{(t)}}, \text{ for } k = 1, \cdots, q. \tag{5.2.5}$$

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 $t = 0$ is 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 stops when the difference between two consecutive Alpha values $\hat{\alpha}^{(t+1)}$ and $\hat{\alpha}^{(t)}$ decreases below a predetermined small number such as 0.001, which represents a threshold below which you consider two coefficients to be identical for all practical purposes.

5.2.3 *Example*

I now want to illustrate the calculation of the AC_1 and α agreement coefficients with a practical example. To compute the α coefficient, Aickin recommends to add a pseudo-count[6] of 1 to the total count of subjects, and to distribute it uniformly among all cells to avoid convergence problems with the iterative algorithm. If your experiment uses 3 categories, your table will have 9 cells. Therefore, distributing a pseudo-count of 1 uniformly across cells increases each cell count by $1/9 = 0.11$ approximately. Cells with no subject would now have 0.11 subject allowing Aickin's algorithm to run smoothly.

Example 5.1

To illustrate the calculation of AC_1 and alpha coefficients, let us consider the reliability data of Table 5.1. This data represents the distribution of human subjects suffering from back pain, by pain type, and observing clinician.

[6] 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.

Table 5.1: 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

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 $\widehat{\kappa}_{1\text{G}} = (0.65 - 0.257725)/(1 - 0.257725) = 0.5285$. As for Aickin's Alpha, after 10 iterations I obtained $\hat{\alpha}_{\text{A}} = (0.65 - 0.4121)/(1 - 0.4121) = 0.4047$, 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.

You can obtain a more detailed account of these calculations by downloading the Excel workbook,

www.agreestat.com/books/cac5/chapter5/chapter5examples.xlsx,

and reviewing the content of the worksheet entitled "Example 5.1.

The Kappa, alpha, and AC_1 statistics of example 5.1 are respectively given by 0.322, 0.405, and 0.529. Kappa represents less than half the magnitude of the percent agreement probability $p_a = 0.65$. This dramatic reduction in the magnitude of the percent agreement is a result of Kappa's unduly high chance-agreement correction. AC_1 on the other hand represents more than 80% of the value of the percent agreement, because of a less severe correction for chance agreement. I will explain in the next few sections that Aickin's alpha coefficient measures a dimension of raters' agreement that is different from what Kappa and AC_1 measure. Therefore, a direct comparison between Aickin's alpha and other coefficients may 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[7] based on all subjects. Because the reference population for evaluating alpha is bigger, $\hat{\alpha}_{\text{A}}$ will generally be

[7] A "for-cause" agreement is an agreement situation where both raters classified a subject into the same category for a reason, as opposed to doing it by pure chance.

lower than AC_1 unless the group of subjects does not have those special subjects that may lead to an agreement by chance. Conceptually, Aickin's alpha would be smaller that both Kappa and AC_1. In practice however, Aickin's alpha often exceeds Kappa because of Kappa's excessive chance-agreement correction in some situations.

The next two sections 5.3 and 5.2.1 deal with the theoretical foundations of the alpha and AC_1 statistics and require some limited abstract thinking. Our primary objective in these two 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 subject, and the raters' skill level), how would we evaluate inter-rater reliability?*" This hypothetical situation will lead to the creation of a theoretical framework. But carrying out a real experiment based on a sample of subjects instead of the whole subject population, always results in a loss of information about non-participating subjects. Only the use of special estimation procedures will compensate for these gaps in our knowledge. The result will be a statistical procedure that is susceptible to sampling errors. These errors also known as statistical errors, are discussed in chapter 5 using the techniques of inferential statistics.

Although both sections 5.3 and 5.4 discuss the motivation behind the formulation of AC_1, and that of alpha, they are not essential for using equations 5.2.1 and 5.2.2 in practice with experimental data. Practitioners not interested in this inquiry could skip these two sections without the chapter's readability[8] being affected, and continue with section 5.5 that is devoted to the AC_1 coefficient for multiple raters.

5.3 Aickin's Theory

Aickin (1990) problem was to define a construct α (without a hat) that measures the extent of agreement between two raters A and B in a way that solely reflects the similarities in their knowledge, experience, and judgment. That is α should be insensitive to those agreements that may occur by pure chance, a possibility that cannot be ignored when using discrete measurement scales. Although there could be more than one way of defining such a parameter, Aickin proposed the following definition:

> *"The α parameter is defined as the fraction of the entire subject population made up of subjects that the two raters A and B classified identically for cause, rather than by chance."*

Without providing an explicit definition of the notion of for-cause agreement, Aickin considers that this type of agreement is reached on subjects that are easy to score, also known as textbook subjects in the terminology of Grove et al. (1981).

[8] We nevertheless highly recommend the reading of these two sections for an ind-depth understanding of the concepts.

A subject is considered easy to score when both raters have a strong opinion regarding its membership category. Disagreement as well as chance agreement on the other hand, are assumed to occur only on subjects that are difficult to classify. Any agreement on the hard-to-classify subjects is considered chance agreement. Aickin's theory consists of dividing the target population of subjects into two subpopulations. These are the subpopulation of Hard-To-Score subjects (or H-subjects), and the subpopulation of Easy-to-Score subjects (or E-subjects). Table 5.2 shows an abstract representation of the distribution of N population subjects broken down by rater and subpopulation in a two-level measurement scale study, as Aickin envisioned it.

In Table 5.2, $N_{11}^{(\mathrm{H})}$ for example represents the count of H-subjects in the study population expected to be classified into category 1 by both raters. Likewise, $N_{21}^{(\mathrm{H})}$ is the count of population H-subjects expected to be classified into categories 2 and 1 by raters A and B respectively. More generally, the subject population contains N subjects, $N_{kl}^{(\mathrm{H})}$ of which are expected to be H-subjects and to be classified into categories k and l by raters A and B respectively. $N_{k}^{(\mathrm{E})}$ $(k = 1, 2)$ of the N population subjects are expected to be E-subjects that both raters will classify into the same category k. All cells of Table 5.2 that are colored in black do not contain any population subjects in Aickin's model.

Table 5.2: Distribution of N Population Subjects by Rater, Subpopulation, and Response Category.

Rater A		Rater B				
		Hard Subjects		Easy Subjects		Total
		1	2	1	2	
Hard Subjects	1	$N_{11}^{(\mathrm{H})}$	$N_{12}^{(\mathrm{H})}$	■	■	$N_{1+}^{(\mathrm{H})}$
	2	$N_{21}^{(\mathrm{H})}$	$N_{22}^{(\mathrm{H})}$	■	■	$N_{2+}^{(\mathrm{H})}$
						N_{H}
Easy Subjects	1	■	■	$N_{1}^{(\mathrm{E})}$	0	$N_{1}^{(\mathrm{E})}$
	2	■	■	0	$N_{2}^{(\mathrm{E})}$	$N_{2}^{(\mathrm{E})}$
						N_{E}
Total		$N_{+1}^{(\mathrm{H})}$	$N_{+2}^{(\mathrm{H})}$	$N_{1}^{(\mathrm{E})}$	$N_{2}^{(\mathrm{E})}$	N
		N_{H}		N_{E}		

Aickin's α coefficient is then defined as follows:

$$\alpha = N_{\mathrm{E}}/N, \tag{5.3.1}$$

where $N_{\text{E}} = N_1^{\text{E}} + N_2^{\text{E}}$ is the population count of E-subjects. Note that even if the total count N of population subjects is known, the number N_{E} of population E-subjects will be unknown. After selecting a sample of n subjects for an inter-rater reliability study and rating them, the sub-sample of E-subjects will still be non-identifiable. Therefore, a direct estimation of the α coefficient is not feasible. To solve this problem, Aickin (1990) postulated a statistical model (to be presented in the next subsection) that governs the classification probabilities, and that incorporates α as a model parameter.

Some Remarks about Aickin's Theory

- The raters are assumed to always agree on E-subjects, a disagreement being possible on H-subjects only. Moreover, any such agreement is considered to be for cause.

- Hard-to-score (resp. Easy-to-score) subjects have their hardness (resp. their easiness) intimately tied to both raters' knowledge. Consequently, the configuration of Table 5.2 is specific to the pair of raters being studied, and both raters share the same Hard-to-score, and Easy-to-score subjects.

  This second assumption is the most restrictive in Aickin's theory. In practice two raters are seldom expected to have the same knowledge and skill level. Some subjects that one rater considers hard to score may prove easy to score for another.

5.3.1 *Aickin's Probability Model*

Let the probabilities P_{kl}, $P_{k|\text{H}}^{(\text{A})}$, and $P_{l|\text{H}}^{(\text{B})}$ be defined as follows:

$$\begin{aligned} P_{kl} &= \text{Probability that raters A and B classify a randomly selected subject into categories } k \text{ and } l \text{ respectively,} \\ P_{k|\text{H}}^{(\text{A})} &= \text{Probability that rater A classifies an H-subject into category } k, \\ P_{l|\text{H}}^{(\text{B})} &= \text{Probability that rater B classifies an H-subject into category } l. \end{aligned}$$

In statistical jargon, $P_{k|\text{H}}^{(\text{A})}$ is referred to as the conditional probability that rater A classifies a randomly chosen subject into category k given that the subject was selected from the H-subject sub-population. Aickin's model is based on the following

representation for the probability P_{kl}:

$$P_{kl} = (1-\alpha)P^{(\mathrm{A})}_{k|\mathrm{H}}P^{(\mathrm{B})}_{l|\mathrm{H}} + \alpha\left(\frac{d_{kl}P^{(\mathrm{A})}_{k|\mathrm{H}}P^{(\mathrm{B})}_{l|\mathrm{H}}}{\sum_{k=1}^{q}P^{(\mathrm{A})}_{k|\mathrm{H}}P^{(\mathrm{B})}_{k|\mathrm{H}}}\right), \text{ where } d_{kl} = \begin{cases} 1 & \text{if } k = l, \\ 0 & \text{otherwise.} \end{cases} \quad (5.3.2)$$

This equation can be seen as a direct application of the Bayes' rule[9] in probability theory. The ratio in parentheses represents the conditional probability that both raters classify a subject into the same category k, given that it is an E-subject. In equation 5.3.2, $1-\alpha$ represents the probability of selecting an H-subject and α that of selecting an E-subject. It should also be noted that checking the validity of such a model may not be a simple task.

By multiplying both sides of equation 5.3.2 by d_{kl} and by summing, one obtains the following:

$$\alpha_{\mathrm{A}} = \frac{\sum_{k=1}^{q}P_{kk} - \sum_{k=1}^{q}P^{(\mathrm{A})}_{k|\mathrm{H}}P^{(\mathrm{B})}_{k|\mathrm{H}}}{1 - \sum_{k=1}^{q}P^{(\mathrm{A})}_{k|\mathrm{H}}P^{(\mathrm{B})}_{k|\mathrm{H}}}. \quad (5.3.3)$$

Equation 5.3.3 shows that the α coefficient has a form similar to that of Kappa, with the important exception that the percent chance agreement is computed based on H-subjects only. That is, only the portion of the subject population where the assumption of independence is expected to be satisfied is used to compute the percent chance agreement. This provision surely protects Aickin's coefficient against the paradoxes associated with Kappa.

5.3.2 *Estimating α from a Subject Sample*

Using the n sample subjects that participated in the inter-rater reliability experiment, the α coefficient is estimated by $\hat{\alpha}_{\mathrm{A}}$ defined by equation 5.2.2 where the

[9] The Bayes' rule stipulates that the probabilities $P(F)$ and $P(G)$ of any two events F and G are related as follows: $P(F) = P(G)P(F/G) + (1-P(G))P(F/\overline{G})$, where $\overline{G}$ is the complement event of G and $P(F/G)$ the conditional probability of F given G.

probabilities $P_{k|\mathrm{H}}^{(\mathrm{A})}$ $(k = 1, \cdots, q)$ would be replaced with their sample-based counterparts $p_{k|\mathrm{H}}^{(\mathrm{A})}$. However, a direct calculation of $\hat{\alpha}$ is impossible due to the unknown probabilities $p_{k|\mathrm{H}}^{(\mathrm{A})}$ $(k = 1, \cdots, q)$, and $p_{l|\mathrm{H}}^{(\mathrm{B})}$ $(l = 1, \cdots, q)$. Aickin (1990) proposed to use the maximum likelihood estimates based on the system of three equations defined by equations 5.2.3, 5.2.4, and 5.2.5. The first equation of this system is a version of equation 5.3.3, the second and third equations are obtained by summing both sides of equation 5.3.2 over k and l respectively.

5.4 Gwet's Theory

Unlike Aickin's alpha coefficient, which is defined as the probability that two raters A and B agree for cause, Gwet's AC_1 (see Gwet, 2008a) is defined as the probability that two raters agree given that the subjects being rated are not susceptible to agreement by pure chance. This definition is more in line with the goal set by Cohen (1960) for Kappa. Cohen wanted Kappa to represent "... the proportion of agreement after chance agreement is removed from consideration ..." The two letters A and C in "AC_1 statistic" stand for Agreement Coefficient, while subscript 1 indicates that only total agreement between the two raters (i.e. diagonal elements) is considered as agreement[10]. Another inter-rater reliability coefficient named the AC_2, which considers certain types of disagreements as partial agreements (also referred to as "second-level agreement") is discussed later in this chapter.

A key conceptual difference between AC_1 and alpha lies on the pool of subjects used as basis for computing the coefficients. Aickin's alpha is based on all subjects, while Gwet's AC_1 is based on the sub-population of subjects obtained after removing from the initial population all subjects that may lead to chance agreement. Although Gwet's model has some similarities with Aickin's, the following differences should be mentioned:

- In Aickin's model, any H-subject is hard to score not just for one rater, but for both. Likewise any E-subject will be easy to score for both raters. In Gwet's model, each rater has his/her own group of E-subjects, and his/her own group of H-subjects. Therefore, some of rater A's E-subjects will be rater B's H-subjects, and vice-versa (see Table 5.3).

- In both Gwet's and Aickin's models, any agreement involving an H-subject (with either rater) is by definition considered as agreement by chance. In Gwet's model however, all population H-subjects that would lead to an agreement by

[10] This will also be referred to as first-level agreement throughout the book; hence the use of subscript "1."

chance - assuming they are identifiable - must be removed from the pool of subjects before computing the relative number of for-cause agreement subjects. In Aickin's model, the relative number of for-cause agreement subjects is calculated with respect to the entire subject population.

Table 5.3: Distribution of Population Subjects by Sub-Population of H- and E-Subjects, by Rater, and by Response Category (1,2)

Rater A		Rater B					
		Hard Subjects		Easy Subjects		Total	
		1	2	1	2		
Hard Subjects	1	N_{11}^{HH}	N_{12}^{HH}	N_{11}^{HE}	N_{12}^{HE}	N_{1+}^{H}	N_{HH+}
	2	N_{21}^{HH}	N_{22}^{HH}	N_{21}^{HE}	N_{22}^{HE}	N_{2+}^{HH}	
Easy Subjects	1	N_{11}^{EH}	N_{12}^{EH}	N_{11}^{EE}	0	N_{1+}^{E}	N_{E+}
	2	N_{21}^{EH}	N_{22}^{EH}	0	N_{22}^{EE}	N_{2+}^{E}	
Total		N_{+1}^{H}	N_{+2}^{H}	N_{+1}^{E}	N_{+2}^{E}	N	
		N_{+H}		N_{+E}			

Table 5.3 shows the configuration of the study population of N subjects from which a subject sample will be selected. The quantity N_{12}^{EH} for example, is the count of subjects identified as E-subjects for rater A and as H-subjects for rater B, and expected to be classified into categories 1 and 2 by raters A and B respectively. As previously indicated, subjects identified as E-subjects for both subjects can only lead to an agreement for cause. No disagreement is possible on E-subjects. Hence the two cells with 0 frequency seen in Table 5.3.

Let κ_{1G} be the construct associated with the AC_1 coefficient. It represents the ideal quantity that AC_1 will approximate with the rating data collected from a reliability experiment. If all the information shown in Table 5.3 was known for all subjects of interest (not just those in an experimental sample), then the AC_1 statistic would be free of sampling errors and would be identical to the theoretical construct κ_{1G}, defined for a general number q of categories as follows:

$$\kappa_{1G} = \frac{\sum_{k=1}^{q} N_{kk}^{EE}}{N - \left(\sum_{k=1}^{q} N_{kk}^{HH} + \sum_{k=1}^{q} N_{kk}^{HE} + \sum_{k=1}^{q} N_{kk}^{EH}\right)}, \tag{5.4.1}$$

where the denominator represents the count of population subjects on which no agreement between raters can be reached by pure chance[11]. Note that under the current setting, Aickin's alpha coefficient would be defined as follows:

$$\alpha_A = \frac{1}{N} \sum_{k=1}^{q} N_{kk}^{EE}. \tag{5.4.2}$$

Aickin excludes chance-agreement subjects from the count of agreement subjects in the coefficient's numerator, but does not exclude them from the reference population in the denominator. Therefore, Gwet's AC_1 coefficient, which excludes chance-agreement subjects from consideration entirely is expected to be higher than Aickin's alpha coefficient.

COMPARING κ_{1G} AND α_A

I am not an advocate of Aickin's alpha coefficient for one reason: by excluding subjects that are susceptible to chance agreement from the numerator while leaving them in the denominator, Aickin makes it difficult if not impossible for its coefficient to reach the perfect value of 1. This is particularly the case when "Hard" subjects are present in the subject population. Consequently, Aickin's alpha coefficient could be artificially low for some subject populations.

The rationale that led to equation 5.4.1 can be looked at this way: Cohen (1960) stated an attractive property that he expected his agreement coefficient to satisfy, but ended up formulating Kappa in a way that did not satisfy it. Here is what Cohen (1960) said on page 40 (second paragraph): "The coefficient κ ... is the proportion of agreement after chance agreement is removed from consideration." The denominator in equation 5.4.1 aims at removing chance agreement from consideration before computing the proportion of agreement, by subtracting from the subject population all subjects susceptible to lead to an agreement by pure chance. The formula Cohen ended up developing was rather based on the following false assumption he made on page 38: "A certain amount of agreement is to be expected by chance, which is readily determined by finding the joint probabilities of the marginals." But Kappa does

[11] This denominator may include some disagreements as well as some agreement for cause. Only subjects causing agreement by chance are removed.

not attempt to quantify that "certain amount" of agreement expected by chance. The motive for not doing it is unknown to me. The joint probabilities of the marginals will hardly help quantify chance agreement, which by the way will occur only on an unknown proportion of subjects, and not on all of them.

5.4.1 *The Probabilistic Model*

Although equation 5.4.1 provides a definitional expression for AC_1, it is useless for computing it since "Hard" and "Easy" subjects cannot be identified. What is needed is a probabilistic model that links observed ratings to the theoretical concepts of "Hard" and "Easy" subjects. Although Table 5.3 shows only two response categories for illustration purposes, I assume here that the rater must classify subjects into one of q possible response categories labeled as $k = 1, \cdots, q$.

Let us consider the following events:

- $\mathcal{R}$: The selected subject is an H-subject *(i.e. one of the two raters or both will perform a nondeterministic rating when classifying this subject).*
- A: Both raters A and B agree on the classification of the selected subject.
- $\mathcal{C} = A \cap \mathcal{R}$: Represents an agreement by chance *(i.e. the selected subject is an H-subject, and both raters A and B agree about its classification).*

For any two categories k and l, a straight application of the Bayes' rule reveals that the probability P_{kl} that raters A and B classify a randomly selected subject into categories k and l respectively can be expressed as follows:

$$P_{kl} = P(\mathcal{C})P_{kl|\mathcal{C}} + P\big(\overline{\mathcal{C}}\big)P_{kl|\overline{\mathcal{C}}}, \tag{5.4.3}$$

where $\overline{\mathcal{C}}$ is the event "No Chance Agreement," the complementary event of $\mathcal{C}$, and $P_{kl|\mathcal{C}}$ the conditional probability that raters A and B classify a subject into categories k and l given event $\mathcal{C}$. Since $P(\mathcal{C}) = P(\mathcal{R})P(A|\mathcal{R})$, I propose the following statistical model for the join classification probability:

$$P_{kl} = P(\mathcal{R})\frac{d_{kl}}{q^2} + \big(1 - P(\mathcal{R})/q\big)P_{kl|\overline{\mathcal{C}}}, \tag{5.4.4}$$

where $d_{kl} = 1$ if $k = l$ and $d_{kl} = 0$ if not. Equation 5.4.4 stems from the fact that $P(\mathcal{C})P_{kl|\mathcal{C}} = P(\mathcal{R})P(A/\mathcal{R})P_{kl|\mathcal{C}}$, and from the hypothesis that the probability $P(A/\mathcal{R})$ of agreement given a random rating is $1/q$ and $P_{kl|\mathcal{C}} = d_{kl}/q$. Some authors (e.g. Grove et al., 1981, among others) pointed out that under the assumption of random rating, it may be inappropriate to assign equal probability $1/q$ to all categories.

One of their recommendations was the use of observed marginal probabilities. I do not recommend this and here is why: if a rater believes that one category is more likely than the others to be the correct one, then that category must be selected and the rating process should not even be considered random in the first place. Why would a rater rates H-subjects in the same way E-subjects are?

By multiplying both sides of equation 5.4.4 by d_{kl} and by summing over k and l one obtains:

$$\sum_{k=1}^{q} P_{kk} = P(\mathcal{R})/q + \big(1 - P(\mathcal{R})/q\big)\kappa_{1G}.$$

Consequently κ_{1G} can be expressed as follows:

$$\kappa_{1G} = \frac{P_a - P(\mathcal{R})/q}{1 - P(\mathcal{R})/q}, \text{ where } P_a = \sum_{k=1}^{q} P_{kk}. \tag{5.4.5}$$

To be able to compute κ_{1G} from observed ratings, I need to compute the probability $P(\mathcal{R})$ of random rating, which represents the proportion of Hard-to-Score subjects for all raters combined.

5.4.2 *Quantifying the Probability $P(\mathcal{R})$ of Selecting an H-Subject*

For the sake of simplicity let us start with a simple experiment that involves 2 raters A and B, and 2 categories labeled as 1 and 2. I am also going to assume that the distribution of subjects across categories is the only information at our disposal that tells us how the raters classified the subjects. That is, no information external to the experiment exists on the subjects and the level of difficulty raters may experience in rating them. I want to use the distribution of subjects to quantify the propensity for random rating. In other words, given the observed distribution of subjects by category and by rater, what is the probability that a rater classifies a subject randomly?

The distribution of raters across categories is described by the 2 probabilities π_1 and π_2 with $\pi_2 = 1 - \pi_1$. These classification probabilities are aggregate measures of individual raters' propensity for classification in specific categories (see equation 5.2.1). The use of summary classification probabilities aims at minimizing individual rater effects when quantifying the probability for a subject to be classified randomly. Each pair (π_1, π_2) can be represented by a point on the diagonal line of Figure 5.1. The arbitrary point $X(\pi_1, \pi_2)$ can move from the top at point A down to the bottom at point C. Each position of point X is associated with a value of $P(\mathcal{R})$.

A key assumption that is made here is that point M, which is in the middle between points A and C, is where the probability of random rating is at its maximum[12].

[12]This assumption is based upon the fact that a random assignment of ratings to subjects is

This probability reaches its minimum value at the two extremes A and C. Quantifying $P(\mathcal{R})$ amounts to defining a measure of probability $[0,1] \times [0,1] \longmapsto [0,1]$ that maps any pair (π_1, π_2) to a value in the interval $[0,1]$.

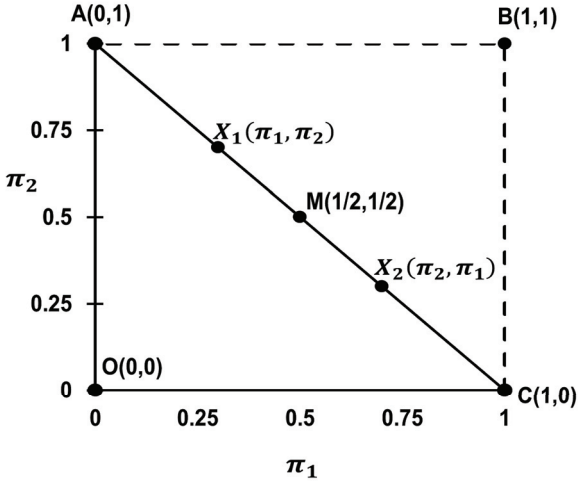

Figure 5.1: Graphical description of the propensity for classification into categories 1 and 2.

Note that any pair (π_1, π_2) can be associated with 2 points $X_1(\pi_1, \pi_2)$ and $X_2(\pi_2, \pi_1)$ on the diagonal line that are symmetric across the middle point M, and which define a rectangle as shown in Figure 5.2. As points X_1 and X_2 move simultaneously towards the 2 extremes A and C, the area of the rectangle expands while the area of the shaded region shrinks. The opposite occurs when the 2 points move towards the center M. The probability of random rating is measured by the area of the shaded region. When the 2 points X_1 and X_2 meet at the center M, the whole rectangle $OABC$ becomes shaded, and its area of 1 represents the maximum value of the probability of radom rating $P(\mathcal{R})$.

One can prove that the area of the shaded region of Figure 5.2 is $4\pi_1(1-\pi_1)$. To facilitate the generalization of the random rating propensity, this expression can also be rewritten as follows:

$$P(\mathcal{R}) = \Big[\pi_1(1-\pi_1) + \pi_2(1-\pi_2)\Big]/(1-1/2),$$

where 2 represents the number of categories q. For an arbitrarily large number of

expected to result in a distribution of subjects described by the probabilities $\pi_1 = \pi_2 = 0.5$ associated with point M.

categories q, $P(\mathcal{R})$ is described as follows:

$$P(\mathcal{R}) = \frac{\sum_{k=1}^{q} \pi_k(1-\pi_k)}{1-1/q}, \tag{5.4.6}$$

where π_k is the probability that a randomly selected subject is classified into category k by a rater (also selected randomly among raters).

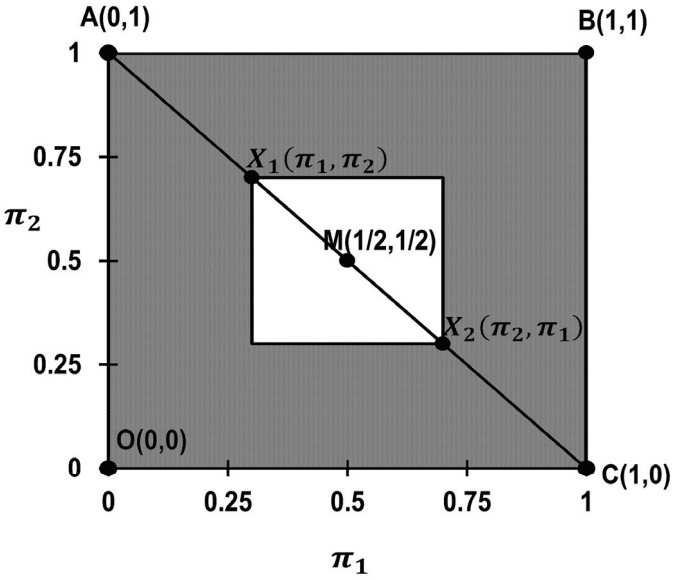

Figure 5.2: Graphical representation of the probability of random rating.

Equation 5.4.6 can also be justified using the chi-square distance between the observed subject distribution and the uniform distribution, from the family of quadratic distances on probabilities whose theoretical foundations were thoroughly studied by Lindsay et al. (2008).. The numerator of equation 5.4.6 would be the observed distance, and the denominator its maximum value.

One way to verify how well equation 5.4.6 works is to simulate a population of subjects similar to what is described in Table 5.3, to define the "true" inter-rater reliability according to equation 5.4.1, and to study the statistical properties of the coefficient given by equation 5.2.1. This verification was done by Gwet (2008a), and the results were very satisfactory.

Formulating κ_{1G} Relative to Model 5.4.4

It follows from equations 5.4.5 and 5.4.6 that the κ_{1G} construct may be rewritten as follows:

$$\kappa_{1G} = \frac{P_a - P_e}{1 - P_e}, \text{ where } P_e = \frac{1}{q-1}\sum_{k=1}^{q}\pi_k(1-\pi_k)\cdot \tag{5.4.7}$$

Equation 5.4.7 shows that the κ_{1G} statistic has a form similar to that of Kappa, with the same percent agreement probability P_a and a different percent chance agreement P_e. To understand how P_e measures chance agreement, let us consider the simple situation where the number of categories is limited to 2. Then $P_e = 2\pi_1(1-\pi_1)$ where π_1 is the probability that classify a subject is classified into category 1. If $\pi_1 = 1$ (i.e. all subjects are classified into category 1), then the percent chance agreement is 0. Intuitively, one can see that if all subjects are systematically classified into one category, then the raters must know what they are doing. An agreement under these conditions is not achieved as a result of pure chance, and the rating process is considered deterministic. On the other hand, if $\pi_1 = 1/2$ (i.e. a randomly selected subject has the same chance to be classified into either category), then $P_e = 0.5$. Again, if the subjects are equally distributed across the categories, then the uniform distribution of subjects matches the configuration that would be obtained if all subjects were H-subjects. The relative number of subjects on the diagonal will then be 50%, which equals P_e.

5.5 Calculating AC_1 for three Raters or More

Section 5.4 introduced the AC_1 coefficient as an abstract construct, the objective being to present an explicit formulation of the concept it represents. This goal was achieved by assuming the hypothetical situation where all subjects of interest as well as their categorization by each of the raters are known. The known includes the raters' knowledge, as well as the group of subjects they consider hard or easy to score. This theoretical framework does not provide the concrete pathway for quantifying the extent of agreement among raters in a practical setting where only observed ratings assigned to subjects are known. Ratings observed during a reliability experiment must be used with valid estimation methods to obtain the concrete value of an agreement coefficient.

An inter-rater reliability experiment is generally based on a sample of n subjects that represents only a fraction of the larger population subjects of interest. The resulting sample-based AC_1 coefficient is denoted by $\widehat{\kappa}_{1G}$ (the hat indicating that it is an approximation of the fixed and unknown abstract κ_{1G} of equation 5.4.1). When the number of raters is limited to two, then equation 5.2.1 is the coefficient that practitioners would use since it provides a good approximation of the population pa-