CHAPTER 2

Agreement Coefficients for Nominal Ratings: A Review

OBJECTIVE

This chapter presents a critical review of several agreement coefficients proposed in the literature in the past few decades for analyzing nominal ratings. Among other coefficients, I discuss the Kappa coefficient of Cohen (1960), its meaning, and its limitations. The different components of Kappa are teased apart and their influence on the agreement coefficient discussed. I explore the case of two raters and two response categories first before expanding to the more general situation of multiple raters and multiple-item response scales. This chapter also treats the important problem of missing ratings often overlooked in the literature. Figure 2.8.1 is a flowchart that shows the different agreement coefficients reviewed, the conditions under which they can be used, and their equation numbers that provide a convenient way to locate them in this chapter.

CONTENTS

"When you can measure what you are speaking about, and express it in numbers, you know something about it. But when you cannot – your knowledge is of meager and unsatisfactory kind. —" .
- Lord Kelvin (1824-1907) -

2.1 The Problem

The objective of this chapter is to present a number of agreement coefficients that have been proposed in the literature for quantifying the extent of agreement among raters when the ratings are data of the nominal type. Such ratings are independent categories, which cannot be ranked neither by order of importance, severity nor any other attribute. Table 2.1 for example shows the distribution of 223 psychiatric patients by diagnosis category and method used to obtain the diagnosis. The first method named "Clinical Diagnosis" (also known as "Facility Diagnosis") is used in a service facility (e.g. public hospital, or a community unit), and does not rely on a rigorous application of research criteria. The second method known as "Research Diagnosis" is based on a strict application of research criteria. Fenning, Craig, Tanenberg-Karant, and Bromet (1994) conducted this study to investigate the extent of agreement between clinical and Research Diagnoses, using the following 4 diagnostic categories:

- Schizophrenia
- Bipolar Disorder
- Depression
- Other

This inter-rater reliability experiment involves two raters, and four possible categories into which the patients may be classified. The two raters are the diagnosis methods "Clinical Diagnosis" and "Research Diagnosis." The rating scale is considered nominal because the four categories cannot be ranked, although it is more accurate to state that this study does not consider the ranking of these categories to be of any interest. The basic problem is to quantify the extent to which the two methods agree about the diagnoses they produce.

The most fundamental and intuitive approach to this problem is to consider the percent agreement as an agreement coefficient. The percent agreement is calculated by summing all four diagonal numbers of Table 2.1 and dividing the sum by the total number of patients. That is,

$$\text{Percent Agreement} = (40 + 25 + 21 + 45)/223 = 131/223 = 58.7\%.$$

Several authors have attempted to identify who first proposed this coefficient. There is still a confusion regarding this issue. While some authors refer to the percent agreement as the Osgood's coefficient, others refer to it as the Holsti's coefficient due

to Osgood(1959) and Holsti (1969) recommending its use at a given point in time. There is ample evidence that the percent agreement has been used numerous times well before these works were released. Since the percent agreement is a crude application of the notion of empirical probability that did not require much investigation, my recommendation would be to put this debate to rest so we can move on to other things.

Researchers observed early in the history of inter-rater reliability estimation that two raters may agree for cause following a clear deterministic rating procedure, or they may agree by pure chance. The problem of chance agreement is best seen in a two-category inter-rater reliability experiment, where two raters must assign subjects to a positive and a negative categories. If two raters are unclear about the categorization of a subject and independently decide to make a subjective choice, they still have a chance to agree that is considerably high given the limited number of options they have to chose from. Because this type of agreement is unpredictable, and difficult to justify, it is clearly not the way any researcher will want the raters to agree. Therefore, agreement by chance is undesirable since it cannot be seen as evidence that the raters master the rating process. Unfortunately the percent agreement accounts for both types of agreement, and can be expected to overstate the "true" extent of agreement among raters. This is the problem that led several authors to propose what is known today as chance-corrected agreement coefficients. The important notion of chance agreement is further discussed in section 2.2.

Psychiatric diagnoses for example, are difficult to make due to the fuzzy boundaries that define various psychiatric disorders. A high degree of consistency between different methods permits each method to validate the other, and eventually be used with confidence and interchangeably on a routine basis. We saw in an example earlier that the clinical and research methods yield the same diagnosis on approximately 58.7% of patients. One can assume that some of these agreements did occur by pure chance. An agreement by chance is not a false agreement. It represents a form of gift or bonus that inflates the relative number of subjects in agreement without resulting from the diagnostic methods' inherent properties. Therefore a patient associated with an agreement by chance does not carry useful information regarding the degree of consistency that can be expected from the methods' intrinsic properties. Consequently, the figure 58.7% overestimates the extent of agreement between the two methods.

If we are able to identify all patients that are susceptible to chance agreement, then we could remove them from our pool of study participants before evaluating the percent agreement. But the sole existence of these special patients does not make them identifiable. A patient is associated with an agreement by chance if the processes that led to a particular diagnosis are not an integral part of the methods.

However, Table 2.1, which constitutes the basis for our analysis, contains no information regarding the processes behind the diagnoses. Moreover, some of these processes may even be cognitive and difficult to capture with precision. Still, an inter-rater reliability coefficient will yield a useful measure of the extent to which two methods are concurrent, only if it is corrected for chance agreement. How one defines chance agreement will determine the form a particular inter-rater reliability coefficient will take.

The oldest chance-corrected agreement coefficient mentioned in the literature is likely from Benini (1901). Other early efforts to solve the chance agreement issue include authors such as Guttman (1945), Bennett et *al.* (1954), Holley and Guilford (1964), Maxwell (1977), Janson and Vegelius (1979), and Brennan-Prediger (1981) who independently developed the same coefficient giving it different names. This simple coefficient, often referred to in the literature as the Brennan-Prediger coefficient is given by the ratio $(p_a - 1/q)/(1 - 1/q)$, where p_a is the percent agreement and q the number of nominal categories in the rating scale.

Brennan and Prediger (1981) recommended their coefficient in the case of two raters and an arbitrary number q of categories, while most authors before recommended it in the simpler case of two raters and two categories only. It can further be extended to the more general case of three raters or more as will be seen in subsequent sections. Holley and Guilford (1964) were the first to formally study this coefficient as a way to compute inter-rater reliability, even though others mentioned it before in various contexts. They named this coefficient the G-Index. These agreement coefficients and many others will be discussed in greater details in subsequent sections.

Some authors have criticized this coefficient under the ground that a practitioner may artificially increase the number of categories. The benefit of this operation would be a smaller chance-agreement probability (i.e. $1/q$), which in turn would increase the magnitude of the agreement coefficient. I believe that this criticism is unfounded. A practitioner who adds dummy categories for the sole purpose of jacking up the agreement coefficient engages in malpractice to obtain an undeserving reward. This is a behavioral problem. Time spent looking for a statistical fix to a problem that stems from a rigged experimental design is not time well spent.

While several inter-rater reliability coefficients have been proposed in the literature since the late forties and early fifties, the Kappa statistic proposed by Cohen (1960) became overtime the most widely-used agreement index of its genre. Despite its popularity, Kappa has many well-documented weaknesses that researchers have been slow to take into consideration when selecting an agreement coefficient. In the next few sections, I will discuss various properties of this coefficient, and will highlight

some of its shortcomings.

Table 2.1: Distribution of 223 Psychiatric Patients by Type of Psychiatric Disorder, and Diagnosis Method

Clinical Diagnosis	Research Diagnosis				Total
	Schizo	Bipolar	Depress	Other	
Schizo	**40**	6	4	15	65
Bipolar	4	**25**	1	5	35
Depress	4	2	**21**	9	36
Other	17	13	12	**45**	87
Total	65	46	38	74	**223**

2.2 Agreement for two Raters and two Categories

A simple inter-rater reliability study consists of evaluating the extent of agreement between two raters who have each classified for example the same 100 individuals into one of two non-overlapping response categories. To be concrete, I will refer to the two raters as A and B and to the two categories as 1 and 2. Ratings obtained from such a study are often organized in a contingency table such as Table 2.2, which contains fictitious data. This table will be used later in this chapter for illustration purposes. Table 2.3 on the other hand, contains similar agreement data in their abstract form. I will appeal to the abstract agreement table throughout this chapter to describe the computational methods in their general form.

Table 2.2 shows that raters A and B both classified 35 of the 100 subjects into category 1, and 40 of the 100 subjects into category 2. Therefore, both raters agreed about the classification of 75 subjects for a percent agreement of 75%. However, they disagreed about the classification of 25 subjects, classifying 5 into categories 2 and 1, and 20 into categories 1 and 2 respectively. Likewise, using the abstract Table 2.3, I would say that raters A and B agreed on the classification of $n_{11} + n_{22}$ subjects out of a total of n subjects for a percent agreement $(n_{11} + n_{22})/n$. If p_a denotes the percent agreement then its value based on Table 2.2 data is given by:

$$p_a = (35 + 40)/100 = 0.75,$$

and its formula given by:

$$p_a = (n_{11} + n_{22})/n\cdot \tag{2.2.1}$$

Table 2.2: Distribution of 100 Subjects by Rater and Category

Rater A	Rater B		Total
	1	2	
1	35	20	55
2	5	40	45
Total	40	60	100

Table 2.3: Distribution of n Subjects by Rater and Category

Rater A	Rater B		Total
	1	2	
1	n_{11}	n_{12}	n_{1+}
2	n_{21}	n_{22}	n_{2+}
Total	n_{+1}	n_{+2}	n

It would seem natural to consider 0.75 as a reasonably high extent of agreement between raters A and B. In reality, this number may overstate what one expects the inter-rater reliability between A and B to be, due to possible chance agreement as discussed in section 2.1. In this section, I will show how Cohen (1960) adjusted p_a for chance agreement to obtain the Kappa coefficient.

CHANCE-AGREEMENT CORRECTION

The idea of adjusting the percent agreement p_a for chance agreement is often controversial, and the definition of what constitutes chance agreement is part of the problem. Rater A for example, ignoring a particular subject's specific characteristics may decide to categorize it randomly[1]. With the number of response categories as small as 2, rater A could still categorize that subject into the exact same group as rater B, creating a lucky agreement that reflects neither the intrinsic properties of the classification system, not rater A's proficiency to use it.

2.2.1 *Cohen's Kappa Definition*

What researchers need, is an approach for measuring agreement beyond chance. To address this problem, Cohen (1960) estimated the expected percent chance agreement (denoted by p_e), and used it to adjust the percent agreement p_a to obtain the Kappa coefficient shown in equation 2.2.3. The percent chance agreement p_e is calculated by summing what Cohen considers to be the two probabilities for the two response categories 1 and 2. Note that the probabilities that raters A and B classify a subject into category 1 are respectively 0.55 and 0.40. These numbers correspond to the raw and column marginal percentages. According to Cohen the two raters are expected to reach agreement on category 1 with probability $0.55 \times 0.40 =$

[1] I consider a subject categorization to be random if it is not based on any known and predetermined process

BAK AND PABAK COEFFICIENTS

Byrt, Bishop, and Carlin (1993) proposed the Bias-Adjusted Kappa (BAK) in an effort to prevent Kappa from producing lower values when table marginals are balanced than when they are unbalanced. The same authors also proposed the Prevalence-Adjusted and Bias-Adjusted Kappa (PABAK) in an attempt to correct for the two paradoxes mentioned by Feinstein and Cicchetti (1990). As it turned out, the BAK statistic is nothing else than the π statistic of Scott (1955), while the PABAK is identical to the Brennan-Prediger statistic discussed earlier in this chapter.

Although Scott's π statistic is not sensitive to the marginal distributions, it remains nevertheless very sensitive to trait prevalence. The Brennan-Prediger statistic on the other hand behaves reasonably well under various conditions. The reader may see a detailed discussion of the merit of these statistics in Gwet (2008*a*).

2.8 Concluding Remarks

The focus of this chapter was the study of various agreement coefficients for measuring inter-rater reliability for two raters, and also to discuss some well-known extensions to the more general case of three raters or more. Only ratings that are of nominal type were considered. These are ratings, which cannot be ranked in any meaningful way and which are generally analyzed as labels that raters assigned to the subjects that are being rated. I provided a detailed discussion of the computational procedures needed to quantify each of the coefficients, and attempted to explain the purpose of each step. In particular, the notion of chance agreement was extensively discussed for all coefficients. While the intuition behind Kappa has always been a good one, I believe that it did not translate well into a formal equation. Cohen's formulation of Kappa has been proven flawed on numerous occasions as can be seen with the different paradoxes described in the literature. Nevertheless, this coefficient remains popular among researchers in the medical and social science fields. I do believe that it is about time for researchers in various fields to give full consideration to some alternative procedures that have been recently proposed in the literature, and which I briefly presented in this chapter, and will further discuss in subsequent chapters.

Also addressed in this chapter is the important and complex problem of missing ratings, which has inexplicably often been overlooked in the literature. I recommended an approach that uses all subjects rated by one rater or more. All subjects not rated at all must be excluded from the analysis. If the number of missing values is high then ignoring subjects that were not rated by all raters would result in a colossal waste of useful information, and possibly a substantial loss of precision in the

calculations. This must be avoided. However, if the number of missing values is small then their impact on the different agreement coefficients is expected to be minimal.

Figure 2.8.1 shows a flowchart with the different agreement coefficients discussed in this chapter, their equation numbers, and the conditions under which they should be used. This may help practitioners quickly identify the specific agreement coefficient they want to use. These agreement coefficients were formulated differently according as the number of raters involved is two or three and more. This distinction was motivated by the different ways ratings are organized when the number of raters is two, and when it is three or more. Therefore, when looking for the right equation to use for computing the agreement coefficient, the first step must be to know the number of raters involved in the experiment. If that number is three or more, then you have six agreement coefficients to choose from. An equation number is associated with each coefficient for fast reference. I also added what I considered to be advantages or disadvantages for using each of the coefficients. If that number is two, then knowing what the number of categories used in the experiment would be the next step, before selecting one of the six agreement coefficients. I did not recommend a particular agreement coefficient nor did I conceal my preference for the AC_1 coefficient to be further discussed in chapter 4. Zhao et al. (2013) discuss a comparative study of various chance-corrected agreement coefficients that I found insightful in many aspects. The reader is encouraged to experiment with some of these coefficients and to compare their properties using a dataset one is familiar with.

Chapter 3 is devoted to ordinal and interval ratings, and special techniques for dealing with them will be discussed. Ordinal categories can be ordered from low to high, and disagreements on adjacent categories are often perceived as partial agreements. These partial agreements will be accounted for in the evaluation of inter-rater reliability. Interval data such as Temperature (e.g. 20^0F, 55^0F), or Year (e.g. 2002, 2009) in addition to being ordered from low to high, can produce meaningful intervals (e.g. 2002 to 2009 represents an interval of 7 years, or 20^0F to 55^0F represents a difference of 35^0F in temperature). The agreement coefficients discussed in this chapter will further be generalized in chapter 3 to provide an efficient approach for handling such data.

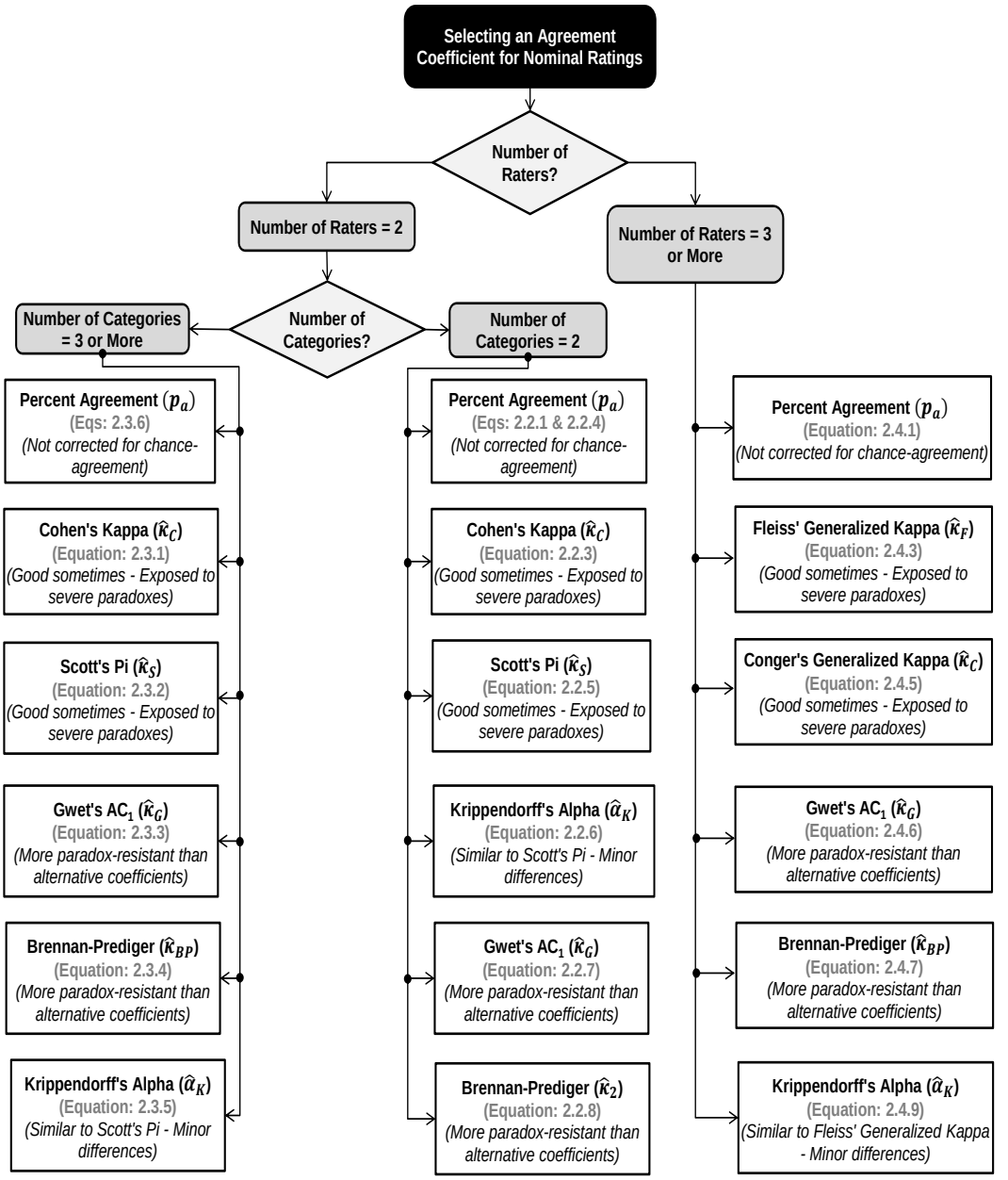

Figure 2.8.1: Choosing an Agreement Coefficient for Nominal Ratings