CHAPTER 4

Agreement Coefficients for Ordinal, Interval and Ratio Data

OBJECTIVE

The objective of this chapter is to extend the study of agreement coefficients to ordinal, interval and ratio data. We will see that the approach recommended by Berry and Mielke Jr. (1988), and Janson and Olsson (2001) for ordinal and interval data reduces to the weighted Kappa proposed by Cohen (1968) when quadratic weights are used. Also extended to ordinal, interval and ratio ratings are Scott's Pi coefficient (Scott, 1955) , Brennan-Prediger statistic (see Brennan and Prediger, 1981), Krippendorff's Alpha coefficient and Gwet's AC_1. These extensions are first described for the simple situation of two raters and two response categories, before being generalized to the case of three raters or more. The little-known generalized Kappa of Conger (1980) is described. Several sets of predefined weights are presented in section 4.6 and provide different ways in which to calibrate partial agreements. Figure 4.9 represents a flowchart showing which agreement coefficient to use (with reference to equation numbers) based on the number of raters and type of ratings.

Contents

4.1 *Overview* . . . 102
4.2 *Generalized Kappa for Two Raters* . . . 103
4.2.1 *Calculating the Kappa Coefficient* . . . 105
4.2.2 *Kappa: a Function of Squared Euclidean Distances* . . . 106
4.3 *Agreement Coefficients for Interval Data: $2 \times q$ Tables* . . . 110
4.4 *Agreement Coefficients for Interval Data and Multiple Raters* . . . 114
4.4.1 *Defining the Multiple-Rater Agreement Coefficient* . . . 115
4.4.2 *Formulating the Multiple-Rater Agreement Coefficient* . . . 116
4.5 *On the Use of Weights for Defining Agreement* . . . 121
4.5.1 *Defining Agreement When Two Measurement Scales Are Used* . . . 121
4.5.2 *Defining Agreement When Raters Assign Some Subjects to Multiple Categories* . . . 125

4.6 *More Weighting Options for Agreement Coefficients* 127
4.7 *Concluding Remarks* 134

"— Without data, you're just another person with an opinion. —" .
- Edward Deming (1900-1993) -

4.1 Overview

Cohen's Kappa coefficient discussed in chapter 2 is suitable only for the analysis of nominal ratings. With nominal ratings, raters classify subjects into categories that have no order structure. That is, two consecutive nominal categories are considered to be as different as the first and last categories. If categories can be ordered (or ranked) from the "Low" to the "High" ends, then the Kappa coefficient could dramatically understate the extent of agreement among raters. Consider an example where a group of adult men are classified twice into one of the categories "Underweight", "Normal", "Overweight" and "Obese" based on their Body Mass Index (BMI). The men are classified the first time using BMI values that are actually measured (i.e. the "Measured" approach). They are classified for the second time using self-reported BMI values (i.e. the "Self-Reported" approach). The problem is to evaluate the extent of agreement between the "Measured" and the "Self-Reported" approaches. Although Kappa may technically be used to evaluate the extent of agreement between the measured and self-reported approaches, we expect it to yield misleading results. The results will be misleading primarily because Cohen's Kappa treats any disagreement as total disagreement. Most researchers would consider the self-reported and measured approaches to be more in agreement if they categorize a participant into the "Overweight" and "Obese" categories, than if they categorize that same participant into the "Underweight" and "Obese" groups. Because it does not account for partial agreement, Kappa as proposed by Cohen (1960) is inefficient for analyzing ordinal ratings. Cohen (1968) proposed the weighted version of Kappa to fix this problem. But what is needed, is a systematic and logical approach for expanding agreement coefficients to handle ordinal as well as interval and ratio data.

Berry and Mielke Jr. (1988), Janson and Olsson (2001), as well as Janson and Olsson (2004) have proposed important extensions of Kappa to ordinal, interval and ratio data[1]. These extensions even allow for the use of multivariate scores on subjects. While a single score determines the subject category membership, the multivariate score on the other hand is a vector of several scores, each being associated with one of the categories. The magnitude of one score associated with a category commensurate with the subject's likelihood of belonging to that category. Situations where a subject could potentially belong to many categories to some degree are common in practice. For example a patient may show symptoms for multiple diseases. Giving raters the

[1]Note that ordinal data can be ranked but the difference between 2 ordinal numbers may have no meaning. Interval data are ordinal data with the exception that the difference between 2 numbers has a meaning although the ratio of 2 numbers may not. With ratio type data however, all arithmetic operations are possible and are meaningful.

option to classify such a patient into more than one categories could prove convenient in some applications.

This chapter is devoted to the various extensions of several agreement coefficients to ordinal, interval and ratio data[2]. While Berry and Mielke Jr. (1988) deserve credit for being among the first to introduce these ideas, I believe that Janson and Olsson (2001) formulated them with more clarity, in addition to further expanding them to handle missing ratings in Janson and Olsson (2004). Therefore, the current presentation is more in line with Janson and Olsson (2001). The reader will notice that the treatment of missing ratings presented in this chapter is substantially different from that of Janson and Olsson (2004), due to my desire to be more practical.

4.2 Generalizing Kappa in the Context of two Raters

Let us consider a simple inter-rater reliability experiment where two raters A and B must each classify 10 subjects into one of two possible categories $+$ (presence of a trait) and $-$ (absence of a trait). Table 4.1 shows the raw ratings as reported by the raters and illustrates what will later be referred to as the raw representation of rating data. Table 4.2 on the other hand, offers an alternative method of reporting the same data that I refer to as the vector representation of ratings.

Table 4.1: Raw Representation of Rating Data

Subject	Rater A	Rater B
1	$+$	$+$
2	$+$	$+$
3	$+$	$-$
4	$+$	$+$
5	$+$	$-$
6	$-$	$+$
7	$-$	$-$
8	$+$	$+$
9	$-$	$-$
10	$+$	$+$

Vector $(1, 0)$ for example, indicates that the rater has classified the subject into the first category (i.e. "+") and not into the second. For this reliability experiment

[2]We assume here that the list of individual ratings that can be assigned to subjects is defined and known before the beginning of the experiment. Otherwise, you should use the intraclass correlation coefficients of chapters 7 through 10.

Table 4.2: Vector Representation of Rating Data

Subject	Rater A	Rater B	Squared Euclidean Distance
1	$(1,0)$	$(1,0)$	0
2	$(1,0)$	$(1,0)$	0
3	$(1,0)$	$(0,1)$	2
4	$(1,0)$	$(1,0)$	0
5	$(1,0)$	$(0,1)$	2
6	$(0,1)$	$(1,0)$	2
7	$(0,1)$	$(0,1)$	0
8	$(1,0)$	$(1,0)$	0
9	$(0,1)$	$(0,1)$	0
10	$(1,0)$	$(1,0)$	0
Total			6

each vector has 2 elements, one for each of the 2 categories "+" and "-". If a three-category measurement scale is used, then a three-dimensional vector such as $(0,1,0)$ will be associated with the raters and the subjects they classified into category 2. With this representation, the rater assigns not a single score to each subject, but rather a *Vector Score (or an Array Score)*. The rightmost column of Table 4.2 represents the discrepancy between rater A's and rater B's ratings measured by the Euclidean distance defined in the next paragraph.

THE EUCLIDEAN DISTANCE

Quantifying how far apart two vectors such as $(1,0)$ and $(0,1)$ are, has traditionally been accomplished with the Euclidean distance defined as $\sqrt{(0-1)^2+(1-0)^2} = \sqrt{2}$. That is, the two vectors are $\sqrt{2}$ units apart. For two arbitrary vectors (a,b) and (c,d), the squared Euclidean distance is given by: $(c-a)^2+(d-b)^2$. This definition indicates that two identical vectors have a distance of 0, which represents a perfect agreement between two raters when vector scores are used. The last column of Table 4.2 contains the squared Euclidean distances between the vector ratings associated with raters A and B. If an inter-rater reliability coefficient is expressed in the form of distances between the raters' respective vector ratings, then generalizing it to ordinal, interval or ratio data will be done in a natural way. This will be feasible since the Euclidean distance has always been used with interval and ratio data.

4.2.1 *Calculating the Kappa Coefficient*

Table 4.3 contains the distribution of the 10 subjects of Table 4.1 by rater and category. From this contingency table and from equations 3.2.1, 3.2.2 and 3.2.3 of chapter 3, it follows that the percent agreement is $p_a = (5 + 2)/10 = 0.7$, while the percent chance agreement is $p_e = (6 \times 7 + 4 \times 3)/100 = 0.42 + 0.12 = 0.54$. Consequently, Cohen's Kappa for raters A and B is given by:

$$\widehat{\kappa}_{\rm C} = \frac{p_a - p_e}{1 - p_e} = (0.70 - 0.54)/(1 - 0.54) = 0.35.$$

Note that Kappa can alternatively be obtained as follows:

$$\begin{aligned}\widehat{\kappa}_{\rm C} &= 1 - \frac{\text{Average of the 10 squared distances of Table 4.2}}{\text{Average of the 100 squared distances of Table 4.4}}, \\ &= 1 - \frac{6/10}{92/100} = 1 - \frac{0.60}{0.92} = \frac{0.32}{0.92} = 0.35.\end{aligned} \tag{4.2.1}$$

To create Table 4.4, each of the 10 vector scores of rater A (c.f. Table 4.2) must be paired with all 10 vector scores of rater B. This pairing process produces 100 pairs of vector scores from both raters. The squared Euclidean distance between the two vector scores of each pair is used to populate Table 4.4.

Equation 4.2.1 shows that Cohen's Kappa is also a function of the squared Euclidean distances between the vector scores of raters A and B. The fact that the Euclidean distance can be used with various data types paves the way for an extension of Kappa to ordinal, interval, or even ratio data.

Table 4.3: Distribution of 10 subjects by Rater and Category

Rater B	Rater A		Total
	+	−	
+	5	1	6
−	2	2	4
Total	7	3	10

Table 4.4: Squared Euclidean Distances between Rater A and Rater B's Vector Scores.

Rater A	Rater B										Total
	$(1,0)$	$(1,0)$	$(0,1)$	$(1,0)$	$(0,1)$	$(1,0)$	$(0,1)$	$(1,0)$	$(0,1)$	$(1,0)$	
$(1,0)$	0	0	2	0	2	0	2	0	2	0	8
$(1,0)$	0	0	2	0	2	0	2	0	2	0	8
$(1,0)$	0	0	2	0	2	0	2	0	2	0	8
$(1,0)$	0	0	2	0	2	0	2	0	2	0	8
$(1,0)$	0	0	2	0	2	0	2	0	2	0	8
$(0,1)$	2	2	0	2	0	2	0	2	0	2	12
$(0,1)$	2	2	0	2	0	2	0	2	0	2	12
$(1,0)$	0	0	2	0	2	0	2	0	2	0	8
$(0,1)$	2	2	0	2	0	2	0	2	0	2	12
$(1,0)$	0	0	2	0	2	0	2	0	2	0	8
Total	6	6	14	6	14	6	14	6	14	6	92

4.2.2 *Kappa*: *a Function of Squared Euclidean Distances*

Let us consider a situation where two raters A and B must classify n subjects into one of two categories 1 or 2. Classifying a subject into one of these categories is equivalent to assigning a two-element vector to that subject. Let A_{ik} be a binary variable, which takes value 1 if rater A classifies subject i into category k (k could be 1 or 2) and will take value 0 otherwise. For rater A, categorizing subject i amounts to assigning a vector score (A_{i1}, A_{i2}) to i. This vector score will be labeled as A_i. Similarly, one can define vector $B_i = (B_{i1}, B_{i2})$ associated with rater B. The Kappa coefficient can then be represented as follows:

$$\widehat{\kappa}_{\mathrm{C}} = 1 - \frac{\frac{1}{n}\sum_{i=1}^{n} d^2(A_i, B_i)}{\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} d^2(A_i, B_j)}, \tag{4.2.2}$$

where $d^2(A_i, B_j)$ represents the squared Euclidean distance from A_i to B_j. Equation 4.2.2 remains the same even if the number of categories q is greater than 2. In the case of q categories A_i and B_i become q-dimensional vectors. For example, the vector

score associated with rater A and subject i will be given by:

$$A_i = (A_{i1}, \cdots, A_{ik}, \cdots, A_{iq}) \tag{4.2.3}$$

Equation 4.2.2 can still be used even if the raters assign one of three interval-type scores x_1, x_2 and x_3 rather than nominal-type scores. In this case, vector A_i will not be a three-element vector consisting of a single occurrence of 1 and two occurrences of 0. Instead, A_i will take a single value (either x_1 or x_2, or x_3 depending on which one is assigned to subject i). As previously seen, this coefficient is identical to the classical Kappa coefficient of Cohen (1960) if x_1, x_2 and x_3 are simply category labels.

When used with q interval data $(x_1, \cdots, x_k, \cdots, x_q)$, equation 4.2.2 leads to the following Kappa coefficient:

$$\widehat{\kappa}_{\mathrm{C}} = 1 - \frac{\sum_{k,l}^{q} p_{kl}(x_k - x_l)^2}{\sum_{k,l}^{q} p_{k+}p_{+l}(x_k - x_l)^2}, \tag{4.2.4}$$

where p_{kl} is the proportion of subjects to whom raters A and B assigned the scores x_k and x_l respectively, p_{k+} is the proportion of subjects to whom rater A assigned score x_k and p_{+l} the proportion of subjects to whom rater B assigned score x_l. This result is obtained from equation 4.2.2 by replacing the vector score A_i with the single interval score x_l $(l = 1, \cdots, q)$ that rater A assigned to subject i.

When dealing with interval data and a complete dataset with no missing rating, you may use equation 4.2.4 for calculating the Kappa coefficient. However, datasets in practice are often incomplete with some raters not rating all subjects. In this case, the proportions p_{kl}, p_{k+}, p_{+l} must be evaluated with respect to the appropriate number of scoring raters. For example, the proportion of subjects that raters A and B classified into categories k and l respectively, must be evaluated only with respect to the subjects that both raters have scored. If a subject was scored by only one of the two raters, then it must be excluded from the calculation of that proportion. Not excluding those subjects will lead to an understatement of the agreement coefficient. Methods for dealing with missing ratings are discussed in section 4.4 and apply to 2 raters as well as to 3 raters or more.

To complete the formulation of the kappa coefficient for ordinal, interval and ratio data, let us introduce the notion of weight. A weight labeled as w_{kl} is associated with any pair of categories k and l and is defined as follows:

$$w_{kl} = \begin{cases} 1 - (x_k - x_l)^2/(x_{max} - x_{min})^2 & \text{, if } k \neq l, \\ 1 & \text{, if } k = l, \end{cases} \tag{4.2.5}$$

where x_{max} and x_{min} are the largest and smallest scores respectively. The purpose of this weight is to quantify how "close" two categories are to one another. The closer the categories, the higher the weight with a maximum value of 1. The particular set of weights described by equation 4.2.5 is known in the literature as "Quadratic Weights, " and additional sets of weights with different weighting goals are presented in section 4.6.

How are these weights calculated when the scores are ordinal and alphabetic such as LOW, MEDIUM, HIGH? The commonly-used approach in this case, is to assign integer values 1, 2 and 3 sequentially to categories according to their ascending order. That is 1, 2 and 3 will be assigned to LOW, MEDIUM and HIGH respectively.

The Kappa coefficient for interval data when the number of raters is limited to two, has the following general form:

$$\widehat{\kappa}_{\mathrm{C}} = \frac{p_a - p_e}{1 - p_e}, \text{ where } p_a = \sum_{k,l}^{q} w_{kl} p_{kl}, \text{ and } p_e = \sum_{k,l}^{q} w_{kl} p_{k+} p_{+l}, \tag{4.2.6}$$

Equation 4.2.6 describes what is known in the literature as the weighted Kappa coefficient of Cohen (1968). As suggested by Cohen (1968), when defining the weighted kappa, the researcher may well define a custom set of weights that may best describe the experimental design. Later in this chapter, I will present alternative sets of weights that were proposed in the literature.

Equation 4.2.5 indicates that the weights typically take values between 0 and 1, the maximum value being reached when a category is associated with itself. In Cohen's terminology, the maximum value of 1 represents "full" agreement while any other value that exceeds 0 represents "partial" agreement. A value of 0 will be seen as total absence of agreement. A non-zero weight (i.e. $w_{kl} > 0$) tides two categories k and l together. In this case, I will say that these 2 categories are affiliated and the magnitude of this weight represents the degree of affiliation of l towards k. Any category will always be affiliated with itself to the maximum degree of 1.

To evaluate Cohen's kappa in practice, the percent agreement p_a should be solely based on subjects that were rated by both raters. Any subject rated by a single rater or not rated at all must first be removed from the calculations. Calculating the percent chance agreement p_e on the other hand, requires the marginal percentages p_{k+} and p_{+l} to be evaluated separately in such a way that p_{k+} (resp. p_{+l}) would be based on all subjects that rater 1 (resp. rater 2) has rated wether rater 2 (resp. rater 1) has rated them or not. The following example illustrates the calculation of the weighted and unweighted Kappa coefficients using a dataset that contains missing ratings.

Example 4.1

Consider the rating dataset of Table 4.5 where two raters named Rater1 and Rater2 have classified 11 units into one of the three categories labeled as A, B and C. As it appears some units where rated by only one of the two raters (units not rated by either rater must be excluded from analysis).

Table 4.5: Rating of 11 subjects by 2 Raters

Units	Rater1	Rater2
1	A	
2	B	C
3	C	C
4	C	C
5	B	B
6	B	
7	A	A
8	A	B
9	B	B
10	B	B
11		C

Table 4.6 shows the distribution of units by rater and includes marginal totals and percentages. This summary table is convenient for experiments involving a large number of units or subjects.

Table 4.6: Distribution of Subjects by Rater and Response Category

	Rater 2					
Rater1	A	B	C	Missing	Total	Row %
A	1	1	0	1	3	27.3%
B	0	3	1	1	5	45.5%
C	0	0	2	0	2	18.2%
Missing	0	0	1	0	1	9.1%
Total	1	4	4	2	11	100%
Column %	9.1%	36.4%	36.4%	18.2%	110%	

Table 4.7 on the other hand, shows the quadratic weights associated with the three categories A, B and C. These weights are assigned to the categories under the assumption that the ranking $A \rightarrow B \rightarrow C$ (i.e. C is ranked higher than B, which in turn is ranked higher than A) represents their correct ascending order. It follows from this table that all diagonal elements equal 1 and represent "full agreement," while off-diagonal elements have a weight of 0 or 0.75 representing "partial agreement." To compute these quadratic weights from equation 4.2.5, I initially assigned the numbers 1, 2 and 3 to the three categories A, B and C respectively (note: if the categories are numeric, then these same numeric values must be used to compute the weights).

The weighted Kappa is given by,

$$\widehat{\kappa}'_{\text{C}} = \frac{0.9375 - 0.7194}{1 - 0.7194} = 0.7772.$$

Readers who want more details regarding these calculations may download the Excel workbook,

www.agreestat.com/books/cac5/chapter4/chapter4examples.xlsx.

This Excel file contains the "Example 4.1" worksheet with all the steps leading to the weighted Kappa. Note that if you replace quadratic weights with identity weights defined by a square matrix with all diagonal elements equal 1 and all off-diagonal elements equal 0, then you will obtain the unweighted kappa of chapter 3. The unweighted kappa is given by,

$$\widehat{\kappa}_{\text{C}} = \frac{0.75 - 0.3444}{1 - 0.3444} = 0.61864.$$

Using the "Example 4.1" worksheet of the Excel workbook chapter4examples.xlsx, you can manually substitute all off-diagonal elements (in red) with zeroes and the entire worksheet will automatically be updated showing the unweighted analysis.

Table 4.7: Quadratic Weights for a three-Level Nominal Scale

	A	B	C
A	1	0.75	0
B	0.75	1	0.75
C	0	0.75	1

4.3 Agreement Coefficients for Interval Data: the Case of two Raters

The fact that this section's title refers to 2 raters does not mean that the discussions will be limited to inter-rater reliability experiments with only 2 participating raters. Instead, it means that the discussions focus on experiments when each