CHAPTER 3

Agreement Coefficients for Ordinal and Interval Data

OBJECTIVE

The objective of this chapter is to extend the study of agreement coefficients to ordinal and interval data. We will see that the approach recommended by Berry and Mielke (1988) , and Janson and Olsson (2001) for ordinal and interval data reduces to the weighted Kappa proposed by Cohen (1968) with the Quadratic Weights. Also extended to ordinal and interval data are the Pi agreement coefficient of Scott (1955), as well as the Brennan-Prediger statistic (see Brennan & and Prediger, 1981). These extensions are first proposed for the simple situation of 2 raters and 2 response categories, before being generalized to the case of 3 raters or more.

CONTENTS

3.1 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 response categories that have no order structure. That is, two consecutive nominal categories are considered 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 adults are classified twice into one of the categories "Underweight", "Normal", "Overweight", and "Obese" based on their Body Mass Index (BMI). The adults are classified the first time using BMI values that are actually measured (i.e. the "Measured" approach). The adults are classified again using self-reported BMI values (i.e. the "Self-Reported" approach). The problem is to evaluate the extent of agreement between the "Measured" and "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 an adult participant into the "Overweight" and "Obese" categories, than if they categorize that 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 we need, is a systematic and logical approach for expanding agreement coefficients to handle ordinal as well as interval and ratio data.

Berry and Mielke (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 will be 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 option to classify such a patient into more than one categories could prove convenient in some applications.

[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.

This chapter is devoted to the various extensions of several agreement coefficients to ordinal, interval, and ratio data. We will discuss about these extensions in the univariate as well as in the multivariate cases. While Berry and Mielke (1988) deserve credit for being among the first to introduce these ideas, we believe that Janson and Olsson (2001) formulated them with more precision and 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 our treatment of missing ratings is substantially different from that of Janson and Olsson (2004), as we attempted to introduce a little more clarity in the presentation.

3.2 Generalizing Kappa in the Context of 2 Raters and 2 Categories

Let us consider a simple inter-rater reliability experiment where two raters A and B must each classify all 10 subjects into one of 2 possible response categories $+$ (presence of a trait), and $-$ (absence of a trait). Table 3.1 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 3.2 on the other hand, offers an alternative method of reporting the same data that we refer to as the vector representation of ratings.

Table 3.1:
Raw Representation of Rating Data

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

Table 3.2:
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

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

each vector has 2 elements, one for each of the 2 categories "+" and "-". If a 3-category measurement scale is used, then a 3-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 3.2 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 2 vectors are $\sqrt{2}$ units apart. For 2 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 will have a distance of 0, and this will represent agreement between 2 raters when vector scores are used. The last column of Table 3.2 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 carried out in a natural way. This will be feasible since the Euclidean distance has always been used with interval and ratio data.

3.2.1 *Calculating the Kappa Coefficient*

Table 3.3 contains the distribution of the 10 subjects of Table 3.1 by rater and category. From this contingency table and from equations 2.1, 2.2, and 2.3 of chapter 2, it follows that the overall 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:

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

Note that Kappa can alternatively be obtained as follows:

$$\begin{aligned} \widehat{\kappa}_{\mathrm{C}} &= 1 - \frac{\text{Average of the 10 squared distances of Table 3.2}}{\text{Average of the 100 squared distances of Table 3.4}}, \\ &= 1 - \frac{6/10}{92/100} = 1 - \frac{0.60}{0.92} = \frac{0.32}{0.92} = 0.35. \end{aligned} \tag{3.1}$$

To create Table 3.4, each of the 10 vector scores of rater A (c.f. Table 3.2) 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 3.4.

Equation 3.1 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 3.3:
Distribution of 10 subjects by Rater and Category

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

Table 3.4:
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

3.2.2 *Kappa: a Function of Squared Euclidean Distances*

Let us consider a situation where 2 raters A and B must classify n subjects into one of 2 categories 1 or 2. Classifying a subject into one of these categories is equivalent to assigning a 2-element vector to that subject. Let $s_{ik}^{(\mathrm{A})}$ 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 $\left(s_{i1}^{(\mathrm{A})}, s_{i2}^{(\mathrm{A})}\right)$ to i. This vector score will be labeled as $\boldsymbol{s}_i^{(\mathrm{A})}$. The Kappa coefficient can then be represented as follows:

$$\widehat{\kappa}_{\mathrm{C}} = 1 - \frac{\dfrac{1}{n}\displaystyle\sum_{i=1}^{n} d^2\left(\boldsymbol{s}_i^{(\mathrm{A})}, \boldsymbol{s}_i^{(\mathrm{B})}\right)}{\dfrac{1}{n^2}\displaystyle\sum_{i=1}^{n}\sum_{j=1}^{n} d^2\left(\boldsymbol{s}_i^{(\mathrm{A})}, \boldsymbol{s}_j^{(\mathrm{B})}\right)}, \tag{3.2}$$

where $d^2\left(\boldsymbol{s}_i^{(\mathrm{A})}, \boldsymbol{s}_j^{(\mathrm{B})}\right)$ represents the squared Euclidean distance from $\boldsymbol{s}_i^{(\mathrm{A})}$ to $\boldsymbol{s}_j^{(\mathrm{B})}$. Equation 3.2 will not change if the number of categories q is greater than 2. In the case of q categories $\boldsymbol{s}_i^{(\mathrm{A})}$ and $\boldsymbol{s}_i^{(\mathrm{B})}$ become q-dimensional vectors. For example, the vector score associated with rater A and subject i will be given by:

$$\boldsymbol{s}_i^{(\mathrm{A})} = \left(s_{i1}^{(\mathrm{A})}, \cdots, s_{ik}^{(\mathrm{A})}, \cdots, s_{iq}^{(\mathrm{A})}\right) \tag{3.3}$$

Equation 3.2 can be used if the raters assign one of 3 interval-type scores x_1, x_2 and x_3 rather than nominal-type scores. In this case, vector $\boldsymbol{s}_i^{(\mathrm{A})}$ will not be a 3-element vector consisting of a single occurrence of 1 and two occurrences of 0. Instead, $\boldsymbol{s}_i^{(\mathrm{A})}$ 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 3.2 leads to the following Kappa coefficient:

$$\widehat{\kappa}'_{\mathrm{C}} = 1 - \frac{\displaystyle\sum_{k,l}^{q} p_{kl}(x_k - x_l)^2}{\displaystyle\sum_{k,l}^{q} p_{k+}p_{+l}(x_k - x_l)^2}, \tag{3.4}$$

where p_{kl} is the proportion of subjects to whom rater A assigned score x_k and rater B assigned score x_l, 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 3.2 by replacing the vector score $\boldsymbol{s}_i^{(\mathrm{A})}$ 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, then you may use equation 3.4 for calculating the Kappa coefficient. However, datasets in practice are often incomplete with some raters producing ratings on a

limited number of subjects. Therefore, we are recommending a more general formulation of Kappa, whose objective is to ensure that the different proportions of subjects p_{kl}, p_{k+}, or p_{+l} are evaluated with respect to an appropriate baseline. For example, the proportion of subjects that raters A and B classified into categories k and l respectively, must be evaluated with respect to the subjects that both raters have scored. If a subject was scored by only one of the 2 raters, then it will be excluded from the calculation of that proportion. Not excluding those subjects will lead to an understatement of the Kappa coefficient.

- Let $p'_{kl} = p_{kl}/(1 - p_{\text{X}})$, where p_{X} is the relative number of subjects scored by a single rater (either rater A or rater B, but not both), and p_{kl} the relative number of subjects classified into categories k and l by raters A and B respectively.
- $p'_{k+} = p_{k+}/(1 - p_{\text{X}+})$, where p_{k+} is the relative number of subjects that rater A classified into category k, and $p_{\text{X}+}$ the relative number of subjects that rater A did not score.
- $p'_{+l} = p_{+l}/(1 - p_{+\text{X}})$, where p_{+l} is the relative number of subjects that rater B classified into category l, and $p_{+\text{X}}$ the relative number of subjects that rater B did not score.
- Any 2 categories k and l have a weight w_{kl} associated with them, and defined as follows:

$$w_{kl} = \begin{cases} 1 - (x_k - x_l)^2/(x_{max} - x_{min})^2, & \text{if } k \neq l, \\ 1, & \text{if } k = l, \end{cases} \tag{3.5}$$

  where x_{max} and x_{min} are the largest and smallest scores respectively. The set of weights described by equation 3.5 is known in the literature as "Quadratic Weights."

  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 following 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 2, has the following general form:

$$\widehat{\kappa}'_{\text{C}} = \frac{p_a - p_e}{1 - p_e}, \text{ where } p_a = \sum_{k,l}^{q} w_{kl} p'_{kl}, \text{ and } p_e = \sum_{k,l}^{q} w_{kl} p'_{k+} p'_{+l}, \tag{3.6}$$

Equation 3.6 describes what is known in the literature as the weighted Kappa coefficient of Cohen (1968). As suggested by Cohen (1968), when defining weighted

kappa, the researcher may well define a custom set of weights that may best described the experimental design. Later in this chapter, we will present alternative sets of weights that have been proposed in the literature.

In the next example, we will illustrate the calculation of the weighted and unweighted Kappa coefficients using a dataset that contains missing ratings.

Example 3.1

Consider the rating dataset of Table 3.5 where 2 raters named Rater1 and Rater2 have classified 11 units into one of the 3 categories labeled as A, B, and C. As it appears some units where rated by only one of the 2 raters (units not rated by either rater must be excluded from analysis).

Table 3.5: Rating of 12 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 3.6 shows the distribution of units by rater, and includes marginal totals and percentages. This summary table may prove useful in experiments involving a large number of units or subjects.

Table 3.6: Distribution of Subjects by Rater

	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 3.7 on the other hand, shows the quadratic weights associated with the 3 cate-

gories 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 3.5, we 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/book3/chapter3examples.xlsx`,

which contains the "Example 3.1" worksheet with all the steps leading to the weighted Kappa. Note that if you replace quadratic weights with Identity weights where all diagonal elements equal 1, and all off-diagonal elements equal 0, then you will obtain the unweighted kappa of chapter 2. The unweighted kappa is given by,

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

Table 3.7: Quadratic Weights for a 3-Level Nominal Scale

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

3.3 Generalizing Pi, and BP to Interval Data : The Case of 2 Raters

The purpose of this section is to generalize the Pi coefficient of Scott (1955) as well as the Brennan-Prediger (BP) coefficient (see Brennan & Prediger, 1981) to interval and ratio data, using the same approach discussed in section 3.2. That is, the interval or ratio data are used to calculate the weights as in equation 3.5, which in turn are used in the weighted versions of the Pi and BP coefficients. Let us consider the situation where 2 raters A and B must assign one of q interval-type values $(x_1, \cdots, x_k, \cdots, x_q)$ to each subject.

Let $\widehat{\kappa}'_{\rm S}$ be the weighted Pi coefficient. This coefficient is defined as follows:

$$\widehat{\kappa}'_{\rm S} = \frac{p_a - p_e}{1 - p_e},\ \text{where } p_a = \sum_{k,l}^{q} w_{kl} p'_{kl},\ \text{and } p_e = \sum_{k,l}^{q} w_{kl} \pi'_k \pi'_l, \tag{3.7}$$

where $\pi'_k = (p'_{k+} + p'_{+k})/2$, with p'_{k+} and p'_{+k} being defined as in section 3.2.2. Scott's coefficient is often used by researchers, although it shares the same weaknesses of Kappa. The paradoxes discussed in the previous chapter for Cohen's Kappa, will have a negative impact on Scott's Pi as well.

Let $\widehat{\kappa}'_q$ be the weighted BP coefficient[2]. It is defined as,

$$\widehat{\kappa}'_q = \frac{p_a - p_e}{1 - p_e},\ \text{where } p_a = \sum_{k,l}^{q} w_{kl} p'_{kl},\ \text{and } p_e = \frac{1}{q^2} \sum_{k,l}^{q} w_{kl}. \tag{3.8}$$

The BP coefficient has been extensively discussed by Brennan & Prediger (1981), and is known to be resistant to the paradoxes associated with Kappa and Pi. Its chance-agreement probability p_e may at times slightly overstate the propensity for raters to agree by pure chance (see Gwet 2008*a*). The calculation of these 2 coefficients will be illustrated in the following example.

Example 3.2

Let us consider the rating data of Table 3.5. We previously used that dataset to illustrate the calculation of the weighted kappa, and will now use it to compute the weighted Pi and BP coefficients.

Table 3.8 shows the coefficients Kappa, Pi, BP, and the percent agreement in their unweighted and weighted versions. The weighted versions of these coefficients use the quadratic weights.

- Cohen's Kappa is, $\widehat{\kappa}_{\rm C} = (0.9375 - 0.7194)/(1 - 0.7194) = 0.7772$.
- Scott's Pi is, $\widehat{\kappa}_{\pi} = (0.9375 - 0.7429)/(1 - 0.7429) = 0.7569$
- Brennan-Prediger coefficient is, $\widehat{\kappa}_3 = (0.9375 - 0.6667)/(1 - 0.6667) = 0.8125$

You may find more details regarding these calculations in the "Example 3.2" worksheet of the Excel workbook,

`www.agreestat.com/book3/chapter3examples.xlsx`.

[2]Note that Brennan & Prediger (1981) only defined an unweighted agreement coefficient that can only handle nominal scores, and complete datasets with no missing ratings. We are proposing here an extension to interval data that can also handle missing ratings.

Table 3.8: Unweighted and Weighted Coefficients from Table 3.5 Data

Agreement Coefficient	Unweighted	Weighted
Cohen's Kappa	0.6186	0.7772
Scott's Pi	0.6038	0.7569
Brenann-Prediger	0.625	0.8125
Percent Agreement	0.75	0.9375

So far, we have presented methods for calculating agreement coefficients for ordinal, and interval data when the number of raters is limited to 2. In the next section, we will extend these methods to reliability studies that involve 3 raters or more. The focus will still be on Cohen's Kappa, Scott's Pi, and Brennan-Prediger coefficients.

3.4 Generalizing Kappa, Pi, and BP Coefficients to Interval Data: The Case of 3 Raters or More

Equations (3.6), (3.7) and (3.8) are useful for computing the extent of agreement between 2 raters using interval scores, but cannot provide a global measure of agreement among 3 raters or more. When an arbitrarily large number r of raters must assign one of q possible interval scores to each of the n subjects, a global inter-rater reliability coefficient is necessary and can be obtained as will now be explained.

Defining the Multiple-Rater Agreement Coefficient

A chance-corrected agreement coefficient between 2 raters labeled as g and h generally takes the following form:

$$\widehat{\kappa}(g,h) = \frac{p_a(g,h) - p_e(g,h)}{1 - p_e(g,h)} = 1 - \frac{1 - p_a(g,h)}{1 - p_e(g,h)}, \tag{3.9}$$

which is the expression we previously used to extend Kappa to interval data following the approach of Janson and Olsson (2001). To generalize to 3 raters or more, we first form all possible pairs (g,h) of raters out of the initial set of r raters. There is a total of $r(r-1)/2$ such pairs. Then we average each of the 2 expressions $\big[1 - p_a(g,h)\big]$ and $\big[1 - p_e(g,h)\big]$ across all $r(r-1)/2$ pairs of raters. This operation will produce 2 averages. The generalized coefficient is then obtained by replacing the pair-specific ratio $\big[1 - p_a(g,h)\big]/\big[1 - p_e(g,h)\big]$ in equation 3.9 with the ratio of the 2 averages. This is the approach that Conger (1980) used to obtain the correct generalization of Cohen's Kappa to the case of 3 raters or more.

Consider an experiment involving 3 raters A, B, and C. The number of raters is $r = 3$, and $3 \times (3-1)/2 = 3$ pairs of raters can be formed out of that group of raters. These pairs are (A,B), (A,C), and (B,C). Therefore, we can compute agreement