CHAPTER 3

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 (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 3.5, and provide different ways in which to calibrate partial agreements. Figure 3.6.1 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

3.1 Overview **74**
3.2 Generalizing Kappa in the Context of two Raters **75**
• The Euclidean Distance **76**
3.2.1 Calculating the Kappa Coefficient **76**
3.2.2 Kappa: a Function of Squared Euclidean Distances **78**
3.3 Agreement Coefficients for Interval Data: the Case of two Raters **82**
3.4 Agreement Coefficients for Interval Data: the Case of three Raters or More **84**
• Defining the Multiple-Rater Agreement Coefficient **85**
• Formulating the Multiple-Rater Agreement Coefficient **86**
3.5 More Weighting Options for Agreement Coefficients **91**
3.6 Concluding Remarks **98**

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 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 (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 option to classify such a patient into more than one categories could prove convenient

[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.

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 (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.

3.2 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 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 I 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

[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.

3.6 Concluding Remarks

There are two objectives that I wanted to achieve in this chapter: (1) To present the method of Janson and Olsson (2004) that provides a systematic way to extend any chance-corrected agreement coefficient to analyze ratings that are of ordinal or interval types, (2) To present the weighted version of several agreement coefficients that can handle missing ratings, and can use various sets of predefined and custom weights. The Janson-Olsson method made it possible to obtain weighted versions of Cohen's kappa, Scott's Pi, Conger's kappa, Gwet's AC_1, Brennan-Prediger coefficient, or Krippendorff's alpha.

The notion of weighted agreement coefficient introduced by Cohen (1968) has proved useful in many applications. It is indeed understandable that some "slight" disagreements would reveal a common view that two raters have about a particular subject's condition, while other disagreements may indicate a wide gap in the perception the two raters have about the same subject. Therefore, it became necessary to have a agreement coefficient that can incorporate these different degrees of disagreement to provide a more accurate representation of the "true" extent to which the views of the different raters converge. Weighting in this sense is an essential technique when the ratings present an ordinal structure at the minimum. Cohen (1968) confined himself to the case of two raters, and the Kappa coefficient. I extended the presentation of weighted coefficients in this chapter to the more general situation involving three raters or more.

The different weighted agreement coefficients are presented separately for two-rater and multiple-rater inter-rater reliability experiments. This is done for convenience since input ratings are generally organized differently in both situations. When the number of raters is three or more, all weighted agreement coefficients use the same weighted percent agreement except Krippendorff's alpha, whose weighted percent agreement is based on a slightly different expression. The difference between Krippendorff's alpha and Fleiss' generalized kappa remains negligible.

I also presented several predefined sets of weights that you may use with your agreement coefficients of choice. As previously mentioned, there is no recipe for selecting the optimal set of weights. However, all weights aim at incorporating some partial agreements into the calculation of the agreement coefficient. The extent to which you want partial agreements to impact the agreement coefficient can help decide which weight is appropriate. Perfect agreements are assigned a full weight of 1, and partial agreements a smaller weight. If some partial agreements are near as important as the perfect agreements then these partial agreements may be assigned a weight of 0.9 for example. Total disagreements receive a weight of 0. I presented Figures 3.5.1 to 3.5.6 to show how the different sets of weights treat partial agree-

ments. An examination of these figures may help the researcher decide which set of weights is more appropriate for a particular analysis.

Although the weighted agreement coefficients can be used with ratings of interval and even of ratio type, one should remember that chance-corrected agreement coefficients can be used only when the inter-rater reliability experiment produces a limited number of predetermined ratings. Weighted or not, the chance-corrected agreement coefficients cannot handle ratings that belong to a continuum or are unknown prior to the experiment being conducted. In this particular situation, one needs to use intraclass correlation coefficients that are discussed in the third part of this book.

Figure 3.6.1 shows a flowchart describing the conditions under which different weights and weighted agreement coefficients are used. It appears that knowing the data type associated with your ratings is essential for selecting the correct set of weights, and the correct equation for calculating the agreement coefficient. For example if your ratings are purely ordinal, and cannot be treated as interval or ratio data, then the choice of weights that can be used is limited to ordinal weights. The other weights require some arithmetic operations such as subtraction, or division which may not carry any meaning when performed on ordinal ratings. I did not recommend a particular agreement coefficient, even though I discussed the strengths and weaknesses of many of them, and did not conceal my preference for the AC_2 coefficient. The reader is encouraged to experiment with some of these coefficients and to compare their properties.

Zhao et al. (2013) have attempted to compare various chance-corrected agreement coefficients, and the conditions under which one might be preferred over alternatives. I found this study insightful to some extent, provided one is willing to adopt the framework used for comparing the coefficients. Note that Zhao et al. (2013) have studied exclusively the magnitude of the different indexes, and did not attempt to validate them. They did not ask whether the different coefficients were even measuring what they are supposed to measure. For example they included in their analysis the Perreault & Leigh's coefficient (c.f. Perreault & Leigh - 1989), which is the only index I vehemently reject, because it was obtained following a false mathematical derivation as I mentioned in chapter 1.

Chapter 4 discusses in great details the notions of agreement, and chance agreement, and how they are used to derive two chance-corrected agreement coefficients: (i)Aickin's Alpha, and (ii) Gwet's AC_1. All assumptions and underlying models are discussed at length. The discussions in chapter 4 do not start with a description of computation procedures. Instead, the very concept of inter-rater reliability is defined in a formal way within a particular theoretical framework. Statistical methods are then described showing how the coefficients can be computed using observed ratings.

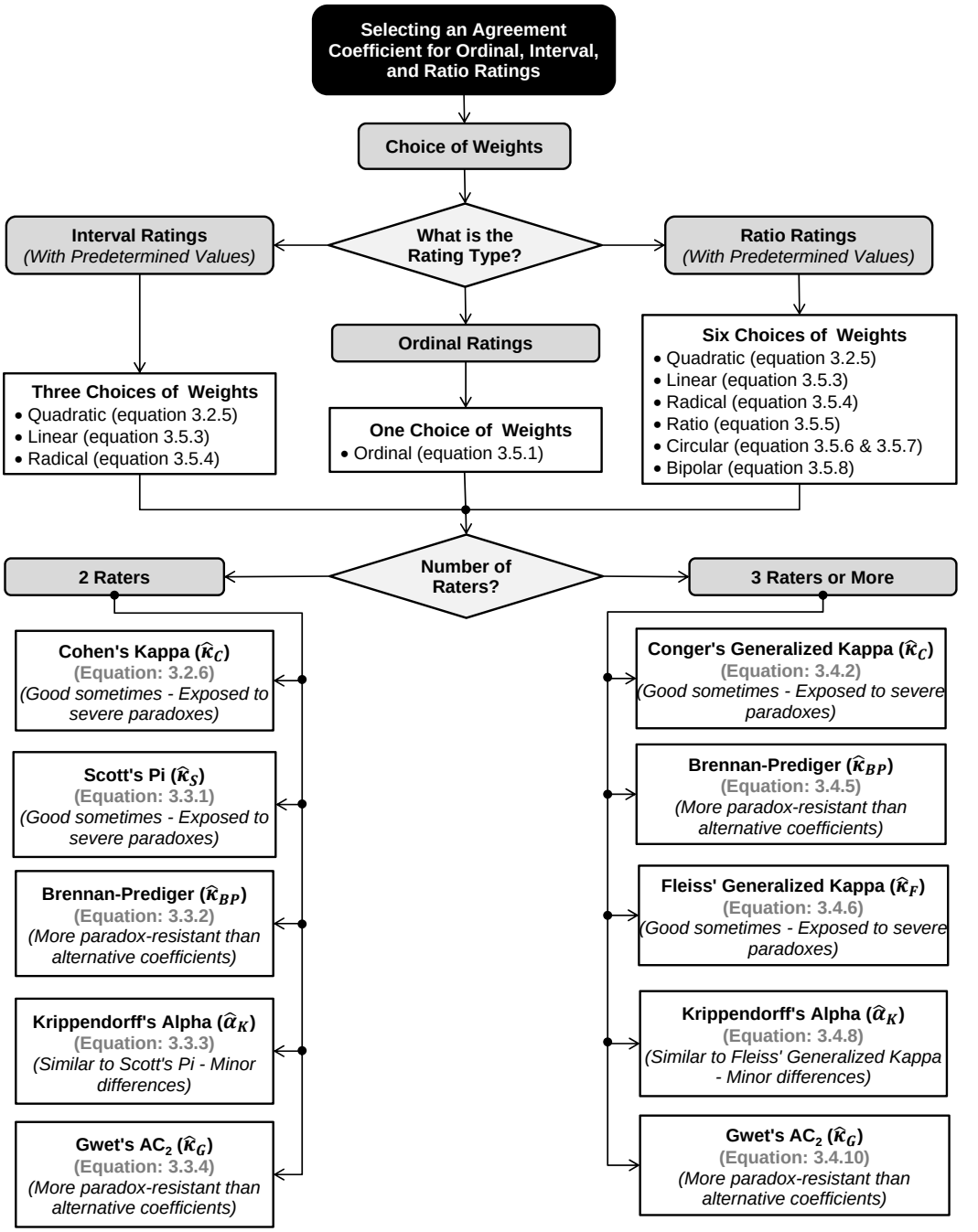

Figure 3.6.1: Choosing a Weighted Agreement Coefficient