CHAPTER 2

The Kappa Coefficient: A Review

OBJECTIVE

This chapter aims at presenting 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. We explore the case of 2 raters and 2 response categories before expanding to the more general situation of multiple raters and multiple-item response scales. This chapter also treats the problem of missing ratings

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

Table 2.1 shows the distribution of 223 psychiatric patients by diagnosis category and method used to make 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

Psychiatric diagnoses 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 on a routine basis. Summing all 4 diagonal numbers of Table 2.1 (i.e. $40+25+21+45 = 131$) gives us an indication of the degree of consistency between the clinical and the research diagnoses. Both methods yield the same diagnosis on approximately 58.7% (obtained as 131/223) of the 223 patients. Researchers have long recognized that in situations similar to this one, some of the 131 agreement patients in Table 2.1 are expected to 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 subject to chance agreement, then we could remove them from our pool of study participants before evaluating the percent of 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.

While several inter-rater reliability coefficients have been proposed in the literature since the 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. In the next few sections, we will discuss various properties of this coefficient, and will illustrate 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 Kappa for 2 Raters and a 2-Level Measurement Scale

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, we will refer to the two raters as A and B and to the two categories as 1 and 2. Classification data obtained from such a study is 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. We will appeal to the abstract agreement table throughout this chapter to describe the computational methods in their general form.

Table 2.2 indicates 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 on the classification of 75 subjects for an overall percent agreement of 75%. However, they disagreed on 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, we 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 an overall percent agreement $(n_{11} + n_{22})/n$. If p_a denotes the overall percent agreement then its value based on Table 2.2 data is given by:

$$p_a = \frac{35 + 40}{100} = 0.75,$$

and its formula is given by:

$$p_a = \frac{n_{11} + n_{22}}{n}. \tag{2.1}$$

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, it overestimates what we expect the inter-rater reliability between A and B to be, due to possible chance agreement as discussed in section 2.1. In this section, we will show how Cohen (1960) adjusted p_a for chance agreement to obtain the Kappa coefficient.

Chance Agreement Correction

The idea of adjusting the overall 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.

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

2.2.1 *Cohen's Kappa Definition*

What researchers need, is an approach to measure agreement beyond chance. To address this problem, Cohen (1960) first estimated the expected percent chance agreement (denoted by p_e), before using it to adjust p_a as shown in equation 2.3 to obtain Kappa. The chance-agreement probability p_e is obtained by summing the two expected agreement probabilities calculated with respect to the 2 response categories 1 and 2. The probabilities for raters A and B to classify a subject into category 1 are respectively 0.55 and 0.40 respectively and represent the raw and column marginal percentages. Therefore the 2 raters are expected to reach agreement on category 1 with probability $0.55 \times 0.40 = 0.22$. Likewise, they are expected to reach agreement on category 2 with probability $0.45 \times 0.60 = 0.27$. Consequently, Cohen's chance-agreement probability is given by :

$$p_e = \frac{55}{100} \times \frac{40}{100} + \frac{45}{100} \times \frac{60}{100} = \frac{49}{100} = 0.49.$$

The chance-agreement probability formula is given by :

$$\begin{aligned} p_e = p_{1+}p_{+1} + p_{2+}p_{+2} &= \frac{n_{1+}}{n} \times \frac{n_{+1}}{n} + \frac{n_{2+}}{n} \times \frac{n_{+2}}{n}, \\ &= p_{1+}p_{+1} + (1 - p_{1+})(1 - p_1), \end{aligned} \tag{2.2}$$

where $n_{1+} = n_{11}+n_{12}$, and $n_{+1} = n_{11}+n_{21}$ are the marginal counts, and p_{1+} and p_{+1} the associated marginal probabilities. Cohen (1960) defined the Kappa coefficient as follows:

$$\boxed{\widehat{\kappa}_{\text{C}} = \frac{p_a - p_e}{1 - p_e}.} \tag{2.3}$$

Although Cohen's original notation for the Kappa was κ (the Greek character "kappa"), we decided to use a different notation $\widehat{\kappa}_{\text{C}}$. In this new notation, the subscript C is the specific label that identifies Cohen's version of Kappa among other versions to be studied, κ_{C} (without the hat) represents the "true value" of Kappa, and the hat ($\widehat{\ }$) indicates an approximation based on sample data. The notion of "true" value reminds us that calculated numbers are always a concrete representation of an abstract (and elusive) reality (some authors will refer to it as a construct) that constitutes our primary interest. These subtleties appeal to more sophisticated statistical concepts in the area of statistical inference, to be discussed in chapter 5.

To understand the meaning of the proposed notation $\widehat{\kappa}_{\text{C}}$, which is further discussed in chapter 5, the reader should remember that the Kappa value is calculated

using one specific sample[2] of subjects. Consequently, a different sample of subjects selected by another researcher is expected to lead to a different value of Kappa. One may then wonder whether there exists a "true", fixed, and unique value for Kappa. The answer is yes, there is a unique "true" Kappa specific to a predefined universe or population of subjects. The subject population of interest is made up of subjects that participated in the reliability study, as well as all those subjects that could potentially be rated in the future and to whom the researcher wants to extend the findings of the inter-rater reliability experiment. Defining this subject population at the beginning of any reliability study is essential for calculating the precision of our statistics.

Kappa's denominator represents the percent of subjects for which one would not expect any agreement by chance, while its numerator according to Cohen (1960) represents *" ... the percent of units in which beyond-chance agreement occurred ..."* Cohen (1960) sees Kappa as a measure of *"... the proportion of agreement after chance agreement is removed from consideration ..."* We will show in chapter 4 that this fundamental goal set by Cohen for Kappa can be achieved with alternative and more efficient methods.

It follows from Table 2.2 data, and from the values of p_a and p_e obtained earlier in this section that the inter-rater reliability between raters A and B as measured by Kappa is given by:

$$\widehat{\kappa}_{\mathrm{C}} = \frac{0.75 - 0.49}{1 - 0.49} \cong 0.51.$$

That is the Kappa-based extent of agreement between raters A and B is approximately equal to 0.51. This represents a "Moderate" agreement level between 2 raters according to the Landis-Koch benchmark scale (see Landis and Koch, 1977). Although widely-used by researchers, this benchmark scale is not without flaws and is further discussed in chapter 6.

2.2.2 *What is Chance Agreement ?*

While the idea of correcting agreement coefficients for chance agreement is justified, the very notion of chance agreement introduced in the previous section is loosely defined. When we claim that 2 raters A and B have agreed by chance, what do we really mean ? Does p_e (Cohen's chance-agreement probability) measure what it is supposed to measure ? These are 2 important questions that need to be addressed.

[2] A sample of subjects in this context does not represent a single unit as is often the case in some medical fields(e.g. a blood sample). Instead, it represents the entire pool of subjects that participated in the reliability study.

- By claiming that raters A and B have agreed by chance in classifying a subject, do we mean that one of the 2 raters not knowing in which category the subject belongs, resolved to take a chance by randomly classifying it (perhaps with an equal probability of 0.5 (i.e. the 50:50 rule)) into one of the 2 possible categories ? This view ties the notion of chance agreement to that of random rating.
- Rather than using the 50:50 rule when randomly categorizing a subject, it is common to consider the marginal classification probabilities p_{1+} and p_{+1} as defining each rater' propensity for classifying a subject into category 1. Even if the rating is random, raters A and B would choose category 1 with probabilities p_{1+} and p_{+1} respectively. They will then agree by chance if one of them performs a random classification according to the observed marginal probabilities. The classification can be seen as having been carried out either independently of the subject's specific characteristics, or following an unknown judgmental process with no apparent logic connecting the subject to the rating.

In both situations described above one of the raters must perform a random classification for concurrence to be considered chance agreement. Based on the second scenario, Cohen (1960) evaluated the chance-agreement probability as shown in equation 2.2. This equation could be problematic for the following reason:

> *The expression* $p_{1+}p_{+1} + (1 - p_{1+})(1 - p_{+1})$ *represents a probability of agreement between raters A and B only if the ratings are known to be independent*[3]*. In case of independence, the overall agreement probability* p_a *and the chance-agreement probability* p_e *will be identical. If the ratings are not independent then the expression* $p_{1+}p_{+1} + (1 - p_{1+})(1 - p_{+1})$ *does not have any particular meaning and does not represent a measure of agreement. Using it in the Kappa equation may yield unpredictable results.*

Krippendorff (2011) argues that Cohen's chance-agreement probability is based on the concept of statistical independence, which in his opinion *"... is only marginally related to how units are coded and data are made and does not yield valid coefficients for assessing the reliability of coding processes ..."*. One of the few instances in statistical science where both expressions p_a ("observed proportion" of agreement) and p_e ("expected proportion" of agreement due to chance) are part of the same equation, occurs when testing the statistical hypothesis of independence between two events

[3]Note that 2 ratings from 2 raters A and B are independent if the knowledge of one makes the other neither more probable nor less probable. This may be the case for a small percent of subjects. If 2 raters have high agreement, then for the majority of subjects, the knowledge of one rating indicates that the other rating is likely to be the same

with the Chi-Square test. In this case, the two expressions are used to define the test statistic, which does not represent any particular metric. Instead, the role of the test statistic is to determine whether the difference between observed and expected values under the hypothesis of independence, is sufficiently large to exclude the possibility that it may have been caused by sampling variation alone. We will further discuss the limitations of Kappa in section 2.5.

2.2.3 *Dealing with Missing Data*

Until now, we have only considered reliability experiments based on fully-crossed designs where each rater must classify all subjects. In practice however, raters may only have the opportunity to classify a portion of the participating subjects. Therefore one rating will be missing for those subjects not rated by both raters. Although the overall percent agreement will be based on the set of subjects rated by both raters, chance-agreement probability on the other hand will use all subjects classified by either rater. Using all subjects will make the marginal probabilities p_{1+} and p_{+1} more precise.

When dealing with missing values, it is convenient to organize the rating data in a contingency table as shown in Table 2.4. Each rater classifies subjects into categories 1 or 2, and all subjects not rated by both raters are classified into a dummy category called X. For example, n_{1X} represents the number of subjects that rater A classified into category 1 and that rater B did not rate at all. Note from Table 2.4 that cell (X, X) always contains a value of 0 (i.e. $n_{XX} = 0$). This indicates that subjects not rated by either rater are excluded from the analysis.

Table 2.4: Distribution of n Subjects by Rater and Response Category with Missing Ratings

Rater A	Rater B			Total
	1	2	X	
1	n_{11}	n_{12}	n_{1X}	n_{1+}
2	n_{21}	n_{22}	n_{2X}	n_{2+}
X	n_{X1}	n_{X2}	0	n_{X+}
Total	n_{+1}	n_{+2}	n_{+X}	n

Considering missing rating data, the overall agreement probability p_a is defined as follows :

$$p_a = \frac{n_{11} + n_{22}}{n - (n_{+X} + n_{X+})}. \tag{2.4}$$

This equation indicates that the baseline for evaluating the overall agreement rate p_a must be restricted to subjects that both raters A and B have rated. Otherwise, the agreement probability will be underestimated since X will be treated as a regular category on which no agreement was achieved.

Cohen's chance-agreement probability on the other hand will still be given by equation 2.2, with the difference that this time $n_{1+} = (n_{11} + n_{12}) + n_{1X}$, and $n_{+1} = (n_{11} + n_{21}) + n_{X1}$ represent the total count of subjects that raters A andB classified into category 1. These counts include subjects rated by a single rater.

Example 2.1

Let us consider a simple and fictitious inter-rater reliability study where 2 raters A and B classified 100 subjects into one of 2 categories labeled as 1 and 2. However, rater B rated 8 subjects that rater A did not have the opportunity to rate. Similarly, rater A rated 5 raters that B did not rate. The ratings are summarized in Table 2.5.

Table 2.5: Distribution of 100 subjects by rater and response category with missing ratings

Rater A	Rater B			Total
	1	2	X	
1	30	18	2	50
2	5	34	3	42
X	5	3	0	8
Total	40	55	5	100

The percent agreement p_a, and percent chance agreement p_e are calculated as follows:

$$p_a = \frac{n_{11} + n_{22}}{n - (n_{X+} + n_{+X})} = \frac{30 + 34}{100 - (5 + 8)} = \frac{64}{87} \approx 0.74.$$

$$p_e = p_{1+}p_{+1} + p_{2+}p_{+2} = \frac{50}{100} \times \frac{40}{100} + \frac{42}{100} \times \frac{55}{100} = 0.431.$$

It follows from equation 2.3 that Kappa is given by:

$$\widehat{\kappa}_C = \frac{0.74 - 0.431}{1 - 0.431} = \frac{0.309}{0.569} \approx 0.54.$$

Using all 100 subjects of example 2.1 to compute the overall percent agreement p_a would reduce it to 0.64 (=64/100) from 0.74. This would be the consequence of considering X as a regular category, and the rating of 13 subjects by only one rater as a disagreement. Ignoring the X category (i.e. missing ratings) will yield a marginal probability[4] p_{+1} of approximately 0.38 ($\approx (30+5)/(50+42)$) as opposed to 0.40 (i.e.

[4]Note that p_{+1} represents the probability for rater B to classify a subject into category 1

40/100) obtained when all 100 subjects are used in the calculation. However, the marginal probability based on 100 subjects is more precise than when it is based on fewer subjects.

Although this section focuses on simple reliability experiments where 2 raters classify subjects into 2 distinct categories, many experiments in practice use more than 2 categories. This generalization is discussed in section 2.3.

2.3 Kappa for 2 Raters on a Multiple-Level Measurement Scale

The Kappa coefficient introduced in section 2.2 within the basic framework of 2 raters and 2 response categories is extended in this section to the more general situation involving 2 raters and an arbitrary number q (greater than 2) of nominal response categories. Such an extension does not present any new conceptual difficulties except when the categories are ordinal instead of nominal. For an ordinal multiple-level scale such as "No", "Possible", "Probable", and "Definite", 2 adjacent categories (e.g. "Probable" and "Definite") although different, still represent a higher degree of agreement[5] than 2 non-adjacent categories (e.g. "No" and "Definite"). Therefore using the Kappa coefficient of section 2.2 with ordinal scales will underestimate the extent of agreement among raters. For such scales, the most effective methods will take into consideration the hierarchical nature of ordinal categories. The problem of ordinal measurement scales is addressed briefly in section 2.6 and in greater details in chapter 3. We confine ourselves in this section to the case where the subjects are scored on a pure nominal measurement scale, where the notion of partial agreement does not apply.

Table 2.6 initially from Sim and Wright (2005) contains reliability data on 2 clinicians (1 & 2) who examined 102 individuals suffering from spinal pain and classified them according to their syndrome type (e.g. "Derangement", "Dysfunction", or "Postural"). For example, the same 11 individuals that clinician 1 diagnosed with a dysfunctional syndrome were diagnosed with a Postural syndrome by clinician 2. The measurement scale used in Table 2.6 is nominal as the 3 response categories ($q = 3$) cannot a priori be ranked in any meaningful way. The extent of agreement between clinicians 1 and 2 can be evaluated using the Kappa coefficient of Cohen (1960).

[5] This special type of agreement on different categories has been referred to as "Partial Agreement" in the literature.

Table 2.6:
Ratings of Spinal Pain by Clinician and Syndrome Type

Clinician 1	Clinician 2			
	Derangement Syndrome	Dysfunctional Syndrome	Postural Syndrome	Total
Derangement Syndrome	22	10	2	34
Dysfunctional Syndrome	6	27	11	44
Postural Syndrome	2	5	17	24
Total	30	42	30	102

In general, reliability data involving 2 raters and q categories will be organized as shown in Table 2.7. This table assumes that each rater had categorized all subjects (i.e. there is no missing rating), and is said to be balanced. Unbalanced tables are discussed later in this section.

Diagonal elements $\{n_{11}, \cdots, n_{qq}\}$ represent the counts of subjects classified into the same category by both raters, while the "Total" column and "Total" row respectively represent raters A and B marginal counts. In practice, percentages will sometimes be used in place of counts. For example $p_{kl} = n_{kl}/n$ represents the percentage of subjects classified into category k by rater A and into category l by rater B, while p_{k+} is the percentage of subjects that rater A classified in category k, and p_{+k} the percentage of subjects that rater B classified in category k.

The Kappa statistic associated with Table 2.7 data is given by:

$$\widehat{\kappa}_{\rm C} = \frac{p_a - p_e}{1 - p_e} \text{ where } p_a = \sum_{k=1}^{q} p_{kk}, \text{ and } p_e = \sum_{k=1}^{q} p_{k+}p_{+k}. \tag{2.5}$$

Before Cohen's Kappa, Scott (1955) suggested the π statistic (read PI statistic) given by:

$$\widehat{\kappa}_{\rm S} = \frac{p_a - p_e}{1 - p_e} \text{ where } p_e = \sum_{k=1}^{q} \hat{\pi}_k^2 \text{ with } \hat{\pi}_k = (p_{k+} + p_{+k})/2. \tag{2.6}$$

Along the lines of Fleiss (1971) we define $\hat{\pi}_k$ as the probability that a rater selected randomly, classify a randomly selected subject into category k. The hat symbol above the π_k character indicates that $\hat{\pi}$ is a sample-based estimated value of the "true" (and unknown) probability π_k, and is likely to change from one group of participating subjects to another. Readers more interested in these notions of "estimated values" and "true" (unknown) parameters will find a more detailed discussion in chapter 5.