CHAPTER 6

Agreement Coefficients and Statistical Inference

OBJECTIVE

This chapter describes several approaches for evaluating the precision associated with the inter-rater reliability coefficients of the past few chapters. Although several factors ranging from the misreporting of ratings to deliberate misstatements by some raters, could affect the precision of Kappa, AC_1 or any other agreement coefficient, the focus is placed on the quantification of sampling errors. These errors stem from the discrepancy between the pool of subjects we want our findings to apply to (i.e. the target subject population), and the often smaller group of subjects that actually participated in the inter-rater reliability experiment (i.e. the subject sample). The sampling error is measured in this chapter by the variance of the inter-rater reliability coefficient. The concept of variance will be rigorously defined, and associated computation methods described. Numerous practical examples are presented to illustrate the use of these precision measures.

Contents

Without theory, experience has no meaning, ⋯ Without theory, one has no questions to ask. Hence without theory, there is no learning.
- Edwards Deming (1900-1993) -

6.1 Introduction

While the past few chapters were primarily devoted to computing various agreement coefficients, the present one aims at exploring what can realistically be done with these numbers. What story do they tell? Can we trust them? If yes, to what extent? In this section, I will provide a more detailed description of the problem at hand, the challenge it poses and how to go about it. Subsequent sections deal with the details of the recommended solutions.

6.1.1 *The Problem*

Tables 6.1 and 6.2 are two representations of a hypothetical rating dataset that Conger (1980) used as examples to illustrate the Kappa coefficient. The ratings are those of 4 raters R_1, R_2, R_3, and R_4 each of whom classified 10 subjects into one of 3 possible categories labelled as a, b, and c. Applied to this data, Fleiss' generalized Kappa (see equation 3.4.3 of chapter 3) yields an inter-rater reliability of $\widehat{\kappa}_F = 0.247$. With the number-crunching phase complete, the next step is to uncover the story these numbers are telling about the research problem being investigated. The statistics must be interpreted and the findings presented in the form of actionable information. To interpret the meaning of an agreement coefficient of 0.247 and to understand its real value, the researcher needs to answer some of the following fundamental questions:

- Is 0.247 a valid number? Does it quantify the extent to which the ratings are reproducible? Can the notion of "extent of agreement among raters" be framed with rigor for researchers to have a common understanding of its most important aspects?

- Can we demonstrate the validity of an observed sample-based agreement coefficient by measuring how close it is to a theoretical construct representing the "extent of agreement among raters?"

- A Kappa coefficient of 0.247 is based on a single sample of 10 subjects and 4 raters. Do these 10 participating subjects constitute a large enough sample to prove the reliability of a newly-developed classification system? If Fleiss'generalized Kappa coefficient indeed measures what it is supposed to measure then how accurate is its calculated value of 0.247? Moreover, are the 4 raters in this study

the only ones to use the classification system? How would a different group of raters affect inter-rater reliability?

Table 6.1: Categorization of 10 subjects into 3 groups $\{a, b, c\}$

Subjects	Raters			
	R_1	R_2	R_3	R_4
1	a	a	a	c
2	a	a	b	c
3	a	a	b	c
4	a	a	c	c
5	a	b	a	a
6	b	a	a	a
7	b	b	b	b
8	b	c	b	b
9	c	c	b	b
10	c	c	c	c

Table 6.2: Distribution of 4 Raters by Subject and Category

Subjects	Categories			Total Raters
	a	b	c	
1	3	0	1	4
2	2	1	1	4
3	2	1	1	4
4	2	0	2	4
5	3	1	0	4
6	3	1	0	4
7	0	4	0	4
8	0	3	1	4
9	0	2	2	4
10	0	0	4	4

Asking those questions leads you straight to the field of inferential methods. These methods allow a researcher to use information gathered from the observed portion of the subject universe of interest, and to project findings to the whole universe (including its unobserved portion). Several inferential methods ranging from crude guesswork to the more sophisticated mathematical modeling techniques have been used to tackle real-world problems. The focus in this chapter will be on the methods of statistical inference, which rely on the agreement coefficient's sampling distribution as a gateway to the unexplored universe of all subjects of interest that could not be reached during the inter-rater reliability experiment.

Several authors have stressed out the need to have a sound statistical base for studying inter-rater reliability problems. For example Kraemer (1979), or Kraemer et al. (2002) emphasize the need to use Kappa coefficients to estimate meaningful population characteristics. Likewise, Berry and Mielke Jr. (1988) mentioned the need for every measure of agreement to have a statistical base allowing for the implementation of significance tests. The analysis of inter-rater reliability data has long suffered from the absence of a comprehensive framework for statistical inference since the early works of Scott (1955) and Cohen (1960). This problem stems from the initial and modest goal the pioneers set to confine agreement coefficients to a mere descriptive role. Cohen (1960) saw Kappa as a summary statistic that aggregates rating data into a measure of the extent of agreement among observers who participated

in the reliability study. Variances and standard errors proposed by various authors approximate the variation of agreement coefficients with respect to hypothetical and often unspecified sampling distributions. But without a comprehensive framework for statistical inference, standard errors are difficult to interpret, and hypothesis testing or comparison between different agreement coefficients difficult to implement.

6.1.2 *The Challenge and How to Go About It*

First and foremost, I like to start by making a clear distinction between the abstract mathematical expression that shows the way to compute an agreement coefficient, the concrete value it produces when actual ratings are applied, and the ideal coefficient you could have computed if you knew everything about all subjects and raters of interest[1] (including those that cannot participate in the inter-rater experiment). I will use the terms "Abstract," "Concrete" and "Theoretical" agreement coefficients respectively to make that distinction. The concrete agreement coefficient is a static value such as 0.68, which is the result of blending an abstract coefficient with one specific set of ratings. The abstract coefficient on the other hand can be seen as a procedure, which could lead to one concrete value or the other depending on the set of actual ratings you feed it with. The theoretical agreement coefficient too is a static value just like the concrete coefficient, although the latter is the experimental number we always get, while the former is the unknown and ideal number that we can only dream about. In the jargon of mathematical statistics, the abstract agreement coefficient would be referred to as the estimator, the concrete coefficient as the estimate, and the theoretical coefficient as the estimand (or the parameter). An estimand will often be denoted by κ, its estimator will be $\widehat{\kappa}$ (the hat indicating that it is an approximation), and a particular estimate would be labeled as $\widehat{\widehat{\kappa}}$ (the double hat indicating that a specific set of ratings was applied to the abstract coefficient to obtain a concrete numeric value for the coefficient). A hypothetical value for the true parameter κ will often be denote by κ_0 (the subscript 0 could be replaced with another subscript such as 1, 2, $\cdots$.

The abstract agreement coefficient does not have a specific value that identifies it. Instead, it does have a "sampling distribution" that tells you how often it is expected to exceed any given value or magnitude you can think of. This sampling distribution plays a pivotal role in evaluating the precision of the agreement coefficient, interpreting its magnitude and planning for future studies. Unless you can find a way to obtain this sampling distribution, it is near impossible to address any of the questions presented in section 6.1.1. Without the sampling distribution, you would be limited to a mere contemplation of a concrete coefficient value, with no possibility to properly interpret it or gain further insight into the quality of our numbers.

[1] In the subsequent sections of this chapter I will need to define more formally what it means to know everything about subjects and raters of interest.

Obtaining the sampling distribution associated with a particular (abstract) agreement coefficient of interest is the key challenge that must be overcome. It is only with the sampling distribution that you can link your concrete agreement coefficient and to the "true" theoretical agreement coefficient it estimates and be able to ascertain its validity. I now want to give you a heads-up on how it works. The sampling distribution we will obtain is that of the quantity $Z = (\widehat{\kappa} - \kappa)/\text{SE}(\widehat{\kappa})$, which is often referred to as the Z-score and represents the difference between the estimator and the estimand, standardized by the standard error[2] of the estimator. Although the estimand κ is always unknown, statistical theory still makes it possible to know the sampling distribution of the Z score. Suppose we want to know whether the "true" and unknown coefficient κ exceeds a pre-defined threshold such as $\kappa_0 = 0.70$. You sure do not have the true coefficient κ. All you have is its surrogate $\widehat{\widehat{\kappa}}$ (the concrete agreement coefficient). How do we use the Z score's sampling distribution to answer this question? Consider the following two possibilities:

- If actual ratings lead to a concrete agreement coefficient (or an estimate) $\widehat{\widehat{\kappa}} = 0.65$, then you do not have any data-based evidence that can support your claim that the estimand κ exceeds 0.70. Note that 0.65 would be your best guess of the magnitude of the estimand.

- Suppose that $\widehat{\widehat{\kappa}} = 0.75$, which clearly exceeds the predefined threshold $\kappa_0 = 0.70$. The trouble is that the estimate $\widehat{\widehat{\kappa}}$ is always subject to a statistical error, which may artificially inflate or deflate its magnitude to some extent. Hence the following question: "Can the difference $0.75 - 0.70 = 0.05$ be solely due to the statistical error[3]?" Or perhaps this statistical error is too small to be able to account for this difference[4]. You can get around this difficulty by first letting the "true" agreement coefficient take the hypothetical $\kappa_0 = 0.70$. This assumption will allow you to fully specify the sampling distribution of the Z score (I show later in the chapter how). If the likelihood for the Z score to exceed this difference is very small ("small in this context typically means below 0.05) then one may conclude that the difference D is relatively large despite it being normalized. Consequently the statistical error alone cannot explain the observed difference of 0.05. The only other contributing factor in this difference is the "true" agreement coefficient κ being greater than the hypothetical value $\kappa_0 = 0.70$.

Because the standard error of the abstract agreement coefficient plays a key role in statistical inference as discussed in the previous paragraph, a large portion of

[2]The standard error $\text{SE}(\widehat{\kappa})$ of the estimator is the statistical measure that tells us how far you would expected the agreement coefficient to stray away from its average value.

[3]The statistical error is generally quantified by the standard error of the agreement coefficient estimate. I will show in subsequent sections how the standard error can be calculated

[4]One may see right here why the ratio $D = (\widehat{\widehat{\kappa}} - \kappa_0)/\text{SE}(\widehat{\kappa})$ is essential for addressing this issue.

this chapter is devoted to showing how it can be calculated for various agreement coefficients. Now, I like to further clarify the ultimate goal this chapter aims at. I want to first describe two broad objectives that are often of interest to researchers. The first of these two objectives is out of the scope of this book, while the second will be the main focus.

- The researcher sometimes wants to understand the process by which raters assign subjects to categories. One may want to know what factors affect the classification and to what degree. Here, no particular group of subjects and no particular group of raters is of interest. The only thing that matters is the scoring process. Each individual's score is seen as a sample[5] from the larger set of all possible scores that can be assigned to any particular subject. During a given reliability experiment, each rater may have to provide several scores (or score samples) for different subjects. The score in this context is analyzed in its abstract form with no reference to a particular group of subjects and raters. Although the number of different scores that a rater can assign to a subject (i.e. the size of the population of scores) may be finite, the fact that the analysis does not target any specific group of subjects nor any particular group of raters led statisticians to refer to this approach as "infinite population inference". Infinity for all practical purposes simply means no reference is made to a specific group of subjects or raters, therefore to the number of samples that can be generated. Agresti (1992) recommends this inferential approach that also uses a theoretical statistical model as a way to study the relationship between raters' scores and the factors affecting them. These techniques represent a particular form of statistical inference, but are out of the scope of this book. Readers interested in this problem may also want to look at Shoukri (2010) or von Eye and Mun (2006)

- The framework of inference developed in this chapter assumes that the researcher has a target group of subjects and a target group of raters that are of interest. These two target groups are often larger than what the researcher can afford to include in a reliability experiment. A psychiatrist at a hospital may want the reliability study to only target his group of patients and the group of raters who may be called upon to use a newly-developed diagnosis procedure. If the group of patients is small, the researcher may conduct a census[6] of the patient population, in which case there will be no need for statistical inference since the statistics produced will match the population parameters. If on the other hand, the large size of the patient population could lead to a costly census

[5] A sample (or a score population sample) in this context is a single observation randomly generated by an often unspecified scoring process, which is specific to each rater.

[6] A census refers to the participation of all subjects of interest in the study

that the researcher cannot afford, then a more affordable option is to survey a subgroup of patients. In this case, the results will be projected only to the predefined finite population of patients from which the participating subjects were selected. Note that the same reasoning applies to the population of raters. That is, statistical inference may be required for the subject population, the rater population, or for both populations. This inferential approach is referred to as "Finite Population Inference[7], and will be the focus in this chapter.

6.2 Finite Population Inference in Inter-Rater Reliability Analysis

Let us consider a reliability study that aims at quantifying the extent of agreement among raters with respect to a given scoring method. We assume that R raters form the rater universe named $\mathcal{U}_{\mathcal{R}}$, and are of interest as potential users of the classification method being tested. Likewise, N subjects forming a subject universe named $\mathcal{U}_{\mathcal{S}}$ are of interest after each of them had been identified as a possible candidate to be scored by one of the R raters. The researcher will ideally want to claim that all R raters can rate all N subjects with a high level of agreement. The raters in the rater population of inference, and subjects in the subject population of inference are labeled as follows:

$$\mathcal{U}_{\mathcal{S}} = \{1, \cdots, i, \cdots, N\},$$
$$\mathcal{U}_{\mathcal{R}} = \{1, \cdots, g, \cdots, R\}.$$

Although some of the R raters and some of the N subjects will not participate in the actual reliability experiment, the researcher still wants the experimental results to be applicable to them. One approach for making this feasible is to start by defining inter-rater reliability, the percent agreement and percent chance agreement with respect to these two populations. If the target numbers of subjects (N) and raters (R) are small then all subjects and raters can be included into the reliability experiment at a reasonable cost. If these numbers are large however, the cost of including all raters and subjects of interest into the study will become prohibitive. A solution to this cost problem is often to randomly select a subset of n subjects from the subject population $\mathcal{U}_{\mathcal{S}}$ and another subset of r raters from the rater population $\mathcal{U}_{\mathcal{R}}$. The two subsets referred to as the *"Rater sample"* (denoted by $s_r^\star$) and the *"Subject sample"* (denoted by s_n) define the participants in the inter-rater reliability experiment. In the notation $s_r^\star$, letter s indicates that the group of units (subjects or raters) represents

[7]This framework for statistical inference was invented by a Polish mathematician named Neyman (1934) and is widely used in large-scale social and business survey projects. Key references related to this topic include Cochran (1977), and Särndal et al. (2003). The main advantage of this approach is found in its ability to generate variances that are valid regardless of the sampling distribution of the agreement coefficients.

a sample (not a population), the star ($\star$) indicates that the sample unit is the rater, and r represents the count of raters in the sample. On the other hand s_n (s without the star) represents a sample of n subjects.

Each time an inter-rater reliability experiment is based on a group of subjects or a group of raters that is smaller than the one being targeted, there is a loss of information that will subject resulting agreement coefficients to errors due to sampling (also known as *"Sampling Errors"*). Quantifying this sampling error and using it in all decisions involving the inter-rater reliability, are among the most fundamental goals of statistical inference. If the reliability experiment involves all N subjects and all R raters of interest then no sampling error will be associated with the resulting agreement coefficients, and there will be no need for inference.

6.2.1 *The Notion of Sample*

Some researchers with background in the social or medical sciences tend to refer to each individual subject as a population sample, and to see a group of n subjects as n subject population samples, the same way one would see 10 blood drops in a medical facility as 10 blood samples. However, the selection of an entire group of n subjects as a whole and the selection of an entire rater group as a whole are the most fundamental building blocks in finite population inference. Consequently, the group of n subjects will be referred to as one sample of subjects of size n, while the whole group of r raters will be seen as one sample of raters of size r.

Note that our ultimate goal is to obtain the sampling distribution of the sample-based agreement coefficient. But the magnitude of the agreement coefficient is determined by both the rater and the subject samples, as well as the respective populations they were selected from (see Figure 6.1 - shaded areas represent what is observable in the process). Therefore the sampling distribution of the agreement coefficient is essentially determined by the selection probability of the whole subject and rater samples. Individual subjects and individual raters have a marginal importance in the inference. Only whole samples are relevant.

For the sake of fixing ideas, let us label all r sample raters and all n sample subjects with numbers as follows:

$$s_r^\star = \{1, \cdots, g, \cdots, r\}, \text{ and } s_n = \{1, \cdots, i, \cdots, n\}.$$

The framework of finite population inference requires that both samples s_n and $s_r^\star$ be selected randomly. The random selection of both groups induces a randomization process, which will define the probabilistic structure of statistical inference.

In a target population of N subjects for example, the total number of samples

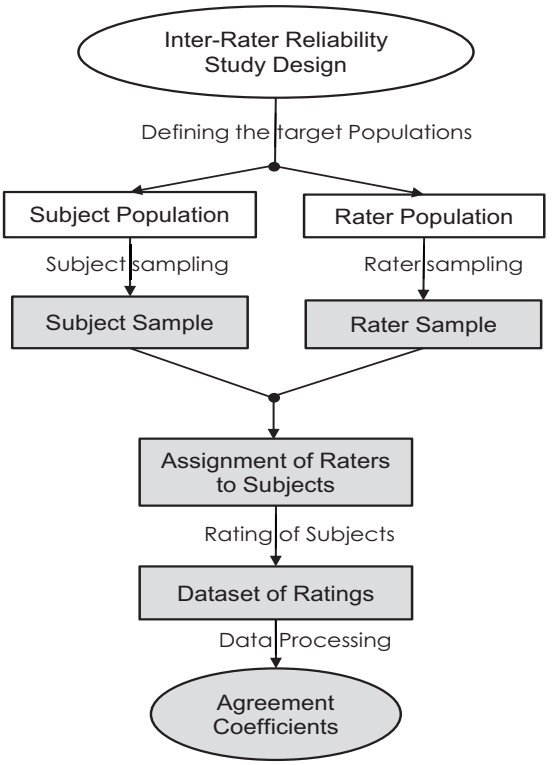

Figure 6.1: The Random Process Leading to the Agreement Coefficients

of size n that one may form is the number of combinations[8] of N objects taken n at a time, and is denoted by $\mathrm{C}_{\mathrm{N}}^{n}$. Likewise $\mathrm{C}_{\mathrm{R}}^{r}$, which is the number of combinations[9] of R raters in groups of r, equals the total count of rater samples of size r one can form from a population of R raters. Note that the researcher will have considerable flexibility in the way the subjects and raters are included in the samples as long as the selection process is random. For example one may decide that all subjects will have an equal chance of being selected for participation in the reliability study, in which case all $\mathrm{C}_{\mathrm{N}}^{n}$ samples of subjects will have the same chance ($p = 1/\mathrm{C}_{\mathrm{N}}^{n}$) of being retained. This is the simple random sampling design. However, the researcher may also decide that one particular subject i_0 has to be part of any participating group for a reason. Such a design will assign a 0 selection probability to all samples not comprising subject i_0. In this case not all samples have the same selection probability. This is

[8]Note that $\mathrm{C}_{\mathrm{N}}^{n} = \binom{N}{n} = N!/\big[n!(N-n)!\big]$ where $N! = N \times (N-1) \times \cdots \times 1$ is N factorial. Moreover, $\mathrm{C}_{\mathrm{N}}^{n}$ can be calculated with MS Excel using the function "$= \mathsf{COMBIN}(N, n)$"

[9]i.e. $\mathrm{C}_{\mathrm{R}}^{r} = \binom{R}{r} = R!/(r!(R-r)!)$

a complex sampling design. In this chapter, we will confine ourselves to the simple random sampling design where all samples have the same selection probability.

6.2.2 *Assigning Raters to Subjects*

The inter-rater reliability study design must include the assignment of raters to subjects. Which rater must rate which subject? There are several design options that are available to the researcher, some of which being more practical and/or efficient than others depending on the type of investigation being conducted. The most commonly-used option in the literature is what I refer to as the *Fully-Crossed* design with no random selection of raters (let us call this the FC_1 design). All raters of interest must rate each subject selected for the inter-rater reliability experiment. A second option is the fully-crossed design with random selection of raters, where only a few raters are randomly selected from an initial pool of raters on interest (let us call this the FC_2 design). In general a fully-crossed design requires the same 3 raters or more to rate all subjects. This may be impractical or expensive or both for various reasons. An attractive alternative that some researchers have used is the *Partially-Crossed* design with 2 raters per subject (let us call this the PC_2 design). According to this design, each subject is assigned 2 raters randomly chosen from an initial target pool of raters.

I like to mention a few scenarios where the researcher might want to consider the PC_2 design. If the rating of human subjects requires a rater-subject interaction that is involved, then it would be unwise to design an inter-rater reliability experiment where 10 raters for example must rate each of the subjects retained. Asking a human subject to accommodate the exact same demanding rating process several times would be unacceptable. The subject needs not be human. It may be a scientific laboratory that is rated by raters working for an accrediting agency. Typically these laboratories must fund the scoring activity performed by an accrediting agency. Accrediting agencies occasionally need to conduct an inter-rater reliability study to ascertain the reliability of the accreditation process. Due to the high cost of accrediting a laboratory, the accrediting agency will want to fund a single rating of the same laboratory in addition to what the lab has funded as part of the regular accreditation process. These two examples shows why the PC_2 design can be useful in some applications.

I will show in subsequent sections of this chapter that for the same number of subjects and raters, the FC_1 design produces agreement coefficients with standard errors that are smaller than those of the PC_2 design. Although the FC_1 design is inconvenient at times, it always produces more accurate agreement coefficients than the PC_2 design when used. Consequently, FC_1 should be the option of choice whenever it is feasible. Unlike the FC_1 design which exposes any agreement coefficient

to up to 2 sources of variation[10], the PC_2 design exposes agreement coefficients to an additional source of variation that stems from the random assignment of pairs of raters to subjects. Nevertheless, the PC_2 design is sometimes the only practical option available to the researcher.

6.2.3 *The Notion of Parameter in Finite Population Inference*

Let i be an arbitrary population subject, and k one of the q response categories into which a rater may classify subject i. If all R raters in the target population were to score subject i, then R_{ik} would be the count of population raters to classify subject i into category k. Likewise $P_{ik} = R_{ik}/R_i$ would be the percent of population raters to classify subject i into category k, where R_i is the count of population raters who rated subject i. The population percent of raters π_k to classify (a subject) in category k is given by:

$$\pi_k = \frac{1}{N}\sum_{i=1}^{N} P_{ik}. \tag{6.2.1}$$

For the sake of clarity and simplicity, let us consider Fleiss' generalized weighted Kappa as an example (Fleiss, 1971). The percent agreement and percent chance agreement, calculated at the population level (i.e. when all target subjects, raters and their respective ratings are known) are respectively denoted by P_a and P_e (the P's are capitalized to indicate that the probabilities are evaluated based on the entire target universe of subjects and raters, and not restricted to samples or subsets of that universe), and defined as follows:

$$P_a = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{q}\frac{R_{ik}(R^{\star}_{ik}-1)}{R_i(R_i-1)}, \quad \text{and } P_e = \sum_{k,l}^{q} w_{kl}\pi_k\pi_l, \tag{6.2.2}$$

where $R^{\star}_{ik}$ is the weighted count of raters who classified subject i into any category affiliated[11] with k (i.e. it is the sum of all $w_{kl}R_{il}$ across all values of l). For a researcher using Fleiss' generalized Kappa coefficient, the parameter of interest κ_{F} for the purpose of inference is defined as,

$$\kappa_{\text{F}} = \big(P_a - P_e\big)\big/\big(1 - P_e\big). \tag{6.2.3}$$

All the quantities P_{ik}, π_k, P_e, P_a or κ_{F} are population parameters to be estimated from the subject and rater samples. We generally use capital Latin letters or Greek

[10] These 2 sources of variation are the selection of subjects and/or the selection raters.

[11] Two categories k and l are affiliated if the corresponding agreement weight w_{kl} takes a non-zero value, and the magnitude of w_{kl} is the degree of affiliation of l towards k. That is a classification of a subject into these categories is seen as partial agreement.

letters for population parameters, while sample-based estimated values of these parameters use small Latin letters, or capital Latin letters with a hat on top. I will formulate these estimated parameters in a way that is unique for the FC_1 and PC_2 designs.

To conduct an inter-rater reliability[12], suppose that a researcher decides to randomly select n out of N subjects of interest, and r out of R raters of interest. Therefore $p_{ik} = r_{ik}/r_i$ is the relative number of times subject i was classified into category k, with r_i being the number of sample raters who provided a rating for subject i. Note that, under the FC_1 design where all r sample raters are expected to rate each subject i, r_i represents the number of ratings associated with subject i and can take integer values ranging from 1 to r, while r_{ik} is the number of times subject i was classified into category k and can take integer values from 0 to r. The situation is slightly different under the PC_2 design. Under this design only 2 of the r sample raters are assigned to rate subject i. Therefore r_i can only take one of the 2 possible values 1, and 2, while r_{ik} can take one of the values 0, 1, and 2. The estimated percent agreement P_a and percent chance agreement P_e respectively denoted by p_a and p_e are defined as follows:

$$p_a = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{q}\frac{r_{ik}(r^{\star}_{ik}-1)}{r_i(r_i-1)}, \quad p_e = \sum_{k,l}^{q} w_{kl}\widehat{\pi}_k\widehat{\pi}_l, \tag{6.2.4}$$

where $r^{\star}_{ik}$ is the weighted number of times that subject i was classified into a category affiliated with k, $\widehat{\pi}_k$ is the estimated propensity for classification into category k. For all practical purposes, it is the relative number of times a subject is classified into category k (or the average of the n values p_{ik} $(i = 1, \cdots, n)$, and represents an estimated value of π_k. For simplicity of notations, we will often use π (without a hat) in place of $\widehat{\pi}$. The estimated value of the Fleiss' agreement coefficient of equation 6.2.3 is denoted by $\widehat{\kappa}_{\mathrm{F}}$ and given by:

$$\widehat{\kappa}_{\mathrm{F}} = \frac{p_a - p_e}{1 - p_e}. \tag{6.2.5}$$

The rater and subject samples must be selected in such a way that the estimated coefficient $\widehat{\kappa}_{\mathrm{F}}$ is as close as possible to its unknown population counterpart κ_{F}. Investigating the relationship between the sample-based $\widehat{\kappa}_{\mathrm{F}}$ and the population-based κ_{F} is a key goal of statistical inference. Note that justifying the particular form that the population-based coefficient takes (e.g. equation 6.2.3)) is not an integral part of the finite population inference framework. This latter task is accomplished with the use of statistical models as shown in chapter 5.

[12] I always assume throughout this book that only subjects rated by one rater or more are considered in the analysis. All subjects that do not receive any rating must be excluded from analysis altogether.

For Gwet's AC_1 or AC_2, the associated population parameters will be respectively κ_{1G} and κ_{2G}. Their sample-based estimates, respectively denoted by $\widehat{\kappa}_{1G}$ and $\widehat{\kappa}_{2G}$, are defined in chapter 5. Their variances are discussed in the next few sections.

6.2.4 *The Nature of Statistical Inference*

Three distinct activities generally define what is known as statistical inference. These are,

- *Point estimation of a population parameter,*
- *Interval estimation of a population parameter,*
- *Test of hypothesis.*

Point estimation is about obtaining a single number as our best approximation of a population parameter using the subject and rater samples. For example p_a will often be our best sample-based approximation of the population parameter P_a. However, the estimation p_a is subject to a sampling error, which may be large. To deal with this error, some researchers will use interval estimation, which provides a range of values in the form of an interval, expected to include the "true" value of the parameter with a high level of confidence. Hypothesis testing on the other hand, determines whether or not a conjecture about the magnitude of a population parameter is consistent with observed ratings. For example the hypothesis that "The Kappa coefficient (at the population level) is greater than 0.20" may or may not be consistent with the ratings observed on a subject sample. Hypothesis testing is a procedure that leads to the rejection or the non rejection of hypotheses.

The inferential procedures of point estimation, interval estimation or hypothesis testing are all built from the sampling distribution of the sample-based agreement coefficients. For example, the overall percent agreement p_a is a function of the subject sample s_n, and the rater sample $s_r^\star$. Consequently, each pair of samples $(s_n, s_r^\star)$ will lead to a different percent agreement value $p_a(s_n, s_r^\star)$. All $C_N^n \times C_R^r$ such pairs of samples lead to a series of $C_N^n \times C_R^r$ values $p_a(s_n, s_r^\star)$, which forms the sampling distribution of p_a upon which statistical inference is built. Expectations, standard errors, and variances are calculated using that discrete sampling distribution. When the subject and rater samples are both generated by a random sampling process, the inference is said to be unconditional . If the subjects are selected randomly and all raters of interest included in the study as participants without sampling, then the inference will be conditional upon the specific group of participating raters. Although not common in practice, the situation where only raters are subject to

random sampling will lead to inference conditionally on the subject sample.

6.2.5 *Independence of Subjects and Its Impact on Statistical Inference*

Some nominal-scale inter-rater reliability experiments require raters to rate each subject more than once. Because the same rater produces several ratings for the same subject, some researchers often conclude that these ratings are no longer independent, and therefore the regular agreement coefficients may no longer be applicable. I like to make a few comments regarding this issue:

- First of all, even when raters assign a single rating to each subject, these ratings are still no independent. This is due to the fact that ratings from different raters are tied to the same subjects, and ratings for different subjects are tied to the same raters.

- There is nothing in the construction of inter-rater reliability coefficients that is based on the assumption of independence. Therefore, computing agreement coefficients will always be technically feasible when rating data is available. In statistical science in general, independence affects the precision with which statistics are calculated, not the calculation of these statistics. They can always be calculated, although they may be more or less depending on the nature of the correlation in sample data.

- For the agreement coefficients discussed in this book, even the precision will not be affected by the correlation between ratings. The variance of nominal-scale inter-rater reliability coefficients is computed using only the sampling distribution induced by the random selection of subjects or the random selection of raters or both. These derivations are design-based and do not rely on any assumptions of independence. The critical thing here is for subjects to be selected randomly. All ratings associated with one subject will have to be treated as if they were produced by different raters for the purpose of computing inter-rater reliability. However, replication can help compute intra-rater reliability as well as inter-rater reliability. This is the main benefit of replication.

6.3 Conditional Inference

This section deals with inferential procedures pertaining to reliability experiments where either the subjects or the raters are selected randomly, but noth both. That is if the subjects participating in the reliability study are selected randomly from the subject population, then no rater other than those participating in the study will be of interest. Likewise, if the participating raters are randomly selected from a larger rater population, then all subjects of interest will be included in the

study. Section 6.3.1 is devoted to the situation where the subject sample is randomly selected from a bigger subject population, but all raters of interest participate in the reliability experiment. Therefore, the statistical error associated with the agreement coefficient will solely be due to the sampling of subjects.

6.3.1 *Inference Conditionally Upon the Rater Sample*

The researcher may decide that only the r participating raters in the rater sample $s_r^\star$ will be of interest, and no effort will be made to project the results beyond that group of raters. The rater sample $s_r^\star$ in this context, is identical to the rater population for the purpose of analysis. Here is a situation where any inter-rater reliability coefficient $\widehat{\kappa}$ will solely be a function of the subject sample. Each subject sample s_n among the $\mathrm{C}_{\mathrm{N}}^n$ possible samples will yield a specific agreement coefficient $\widehat{\kappa}(s_n)$. Therefore, there are $\mathrm{C}_{\mathrm{N}}^n$ possible values for the agreement coefficient, which provide the sampling distribution needed for statistical inference. This inferential procedure will be carried out conditionally upon the specific rater sample $s_r^\star$, and will be referred to as the *Conditional Inference on the Rater Sample* or the *Statistical Inference Conditionally upon the Rater Sample*.

By definition, the "true" or "Population", or "Exact" variance of an agreement coefficient $\widehat{\kappa}$ is the straight variance of all sample-based $\widehat{\kappa}\big(s_n^{(b)}\big)$ values taken on each of the $\mathrm{C}_{\mathrm{N}}^n$ possible subject samples. It is given by:

$$V\big(\widehat{\kappa}(s_n)|s_r^\star\big) = \sum_{b=1}^{\mathrm{C}_{\mathrm{N}}^n} P\Big(s_n^{(b)}\Big)\Big[\widehat{\kappa}\big(s_n^{(b)}\big) - \overline{\widehat{\kappa}}\Big]^2, \tag{6.3.1}$$

where $s_n^{(b)}$ is the b^{th} subject sample, $\overline{\widehat{\kappa}}$ is the average of all $\mathrm{C}_{\mathrm{N}}^n$ possible values that can be taken by the agreement coefficient $\widehat{\kappa}\big(s_n^{(b)}\big)$, and $P\big(s_n^{(b)}\big)$ the probability of selecting the specific sample $s_n^{(b)}$.

Evaluating the variance of an agreement coefficient using equation 6.3.1 is an impossible task. Not only will it be a tedious process to select all possible subject samples of size n out of the target population of N subjects, but implementing equation 6.3.1 would also require each of the N population subjects to have been scored by all raters, which is almost never the case. Consequently the exact variance of the agreement coefficient must be approximated based on a single subject sample and a single rater sample, which is all practitioners have at their disposal. The mathematical formulas used to compute these approximations, are referred to in the statistical literature as *Variance Estimators* as opposed to "Exact" variances such as equation (6.3.1). Gwet (2008a) suggested variance estimators for the AC_1, Kappa, Pi, and Brennan-Prediger (BP) agreement coefficients. These results are summarized

here, and then expanded to accommodate the missing ratings, as well as the use of weights. As in the previous chapters, we will treat two-rater and multiple-rater experiments separately for the sake of clarity.

6.3.1a) VARIANCES FOR TWO RATERS BASED ON CONTINGENCY TABLES

In an inter-rater reliability experiment involving 2 raters A and B (i.e. $r = 2$), the ratings are often summarized as shown in Table 3.5 of chapter 3, where n_{kl} represents the count of subjects that raters A and B classified into categories k and l respectively, and $p_{kl} = n_{kl}/n$ the corresponding percentage. Moreover, $p_{k+} = n_{k+}/n$ and $p_{+k} = n_{+k}/n$ represent raters' A and B marginal classification probabilities respectively. As Fleiss (1971) suggested, $\pi_k = (p_{k+} + p_{+k})/2$ is interpreted as the probability that a randomly selected rater would classify a randomly selected subject into category k. If you are using interval data, then k might represent an interval score x_k instead.

Two-Rater Variances of the AC_1 and AC_2 Statistics

Let $\widehat{\kappa}_{1G}$ and $\widehat{\kappa}_{2G}$ denote the AC_1 and AC_2 statistics respectively. It follows from chapter 5 that both coefficients take the form of a ratio $(p_a - p_e)/(1 - p_e)$ where p_a, and p_e are the percent agreement and percent chance agreement respectively. Assuming that $f = n/N$ is the sampling fraction (i.e. the fraction of the target population that was sampled)[13],

- When there is no missing rating, and the ratings are organized in a contingency table, then the variance of the AC_1 coefficient proposed by Gwet (2008a) is given by:

$$\begin{aligned} v(\widehat{\kappa}_{1G}) &= \frac{1-f}{n(1-p_e)^2}\Bigg\{p_a(1-p_a) - 4(1-\widehat{\kappa}_{1G})\sum_{k=1}^{q} p_{kk}\left(\frac{1-\pi_k}{q-1}\right) \\ &\quad + 4p_e\widehat{\kappa}_{1G}(1-\widehat{\kappa}_{1G}) + 4(1-\widehat{\kappa}_{1G})^2\sum_{k=1}^{q}\sum_{l=1}^{q} p_{kl}\left(\frac{1-\pi_{kl}}{q-1}\right)^2\Bigg\}, \end{aligned} \tag{6.3.2}$$

  where $\pi_{kl} = (\pi_k + \pi_l)/2$.

- When there is no missing rating, and the ratings are organized in a contingency

[13] In many studies the size of the subject population N is unknown, in which case one should set $f = 0$. This amounts to assuming that the sampling fraction is negligible for all practical purposes.

table, then the variance of the AC_2 coefficient (i.e. weighted AC_1) is given by:

$$v(\widehat{\kappa}_{2G}) = \frac{1-f}{n(1-p_e)^2}\left\{\sum_{k,l}^{q} p_{kl}\left[w_{kl} - 2(1-\widehat{\kappa}_{2G})\left(\frac{T_w(1-\pi_{kl})}{q(q-1)}\right)\right]^2 - \left[\widehat{\kappa}_{2G} - p_e(1-\widehat{\kappa}_{2G})\right]^2\right\}, \tag{6.3.3}$$

where T_w is the summation of all agreement weights. Note that equation 6.3.2 is a special case of equation 6.3.3 used with identity set of weights (i.e. weights were all diagonal elements are 1 and all off-diagonal elements are 0).

- If your data contains missing ratings (i.e. some subjects were rated by single rater, as opposed to being rated by both raters), then I recommend to avoid organizing it in a contingency table. Instead, you would have two columns of ratings (one column for each of the two raters), where each row is associated with a subject. The variance expression to used is provided by equation 6.3.17.

Two-Rater Variances of the Unweighted and Weighted Scott's π Coefficient

Scott's π statistic (Scott, 1955) is given by $\widehat{\kappa}_S = (p_a - p_e)/(1-p_e)$, where p_e is Scott's percent chance agreement of equation 3.3.2 in chapter 3.

- When there is no missing rating, and the ratings are organized in a contingency table then the variance of the unweighted Scott's coefficient proposed by Gwet (2008a) is given by:

$$v(\widehat{\kappa}_S) = \frac{1-f}{n(1-p_e)^2}\left\{p_a(1-p_a) - 4(1-\widehat{\kappa}_S)\sum_{k=1}^{q} p_{kk}\pi_k + 4p_e\widehat{\kappa}_S(1-\widehat{\kappa}_S) + 4(1-\widehat{\kappa}_S)^2\sum_{k=1}^{q}\sum_{l=1}^{q} p_{kl}\pi_{kl}^2\right\}, \tag{6.3.4}$$

- When there is no missing rating, and the ratings are organized in a contingency table, then the variance of the weighted Scott's coefficient is given by,

$$v(\widehat{\kappa}_S) = \frac{1-f}{n(1-p_e)^2}\left\{\sum_{k,l}^{q} p_{kl}\left[w_{kl} - 2(1-\widehat{\kappa}_S)\overline{\pi}_{kl}\right]^2 - \left[\widehat{\kappa}_S - p_e(1-\widehat{\kappa}_S)\right]^2\right\}, \tag{6.3.5}$$

where p_e, $\overline{p}_{+k}$, $\overline{p}_{l+}$, $\overline{\pi}_k$, and $\overline{\pi}_{kl}$ are defined as follows:

$$\begin{aligned} p_e &= \sum_{k,l} w_{kl}\pi_k\pi_l,\ \overline{\pi}_k = (\overline{p}_{+k} + \overline{p}_{k+})/2, \\ \overline{p}_{+k} &= \sum_{l=1}^{q} w_{kl}p_{+l},\ \overline{p}_{l+} = \sum_{k=1}^{q} w_{kl}p_{k+},\ \text{and}\ \overline{\pi}_{kl} = (\overline{\pi}_k + \overline{\pi}_l)/2 \end{aligned} \tag{6.3.6}$$

- Your data may contain missing ratings (i.e. some subjects were rated by single rater, as opposed to being rated by both raters). In this case, I recommend organizing it in the form of two columns of ratings[14] instead of organizing it in a contingency table. The variance expression to use is provided by equation 6.3.18.

Two-Rater Variances of the Unweighted and Weighted Cohen's Kappa

The Kappa coefficient (Cohen, 1960) is given by $\widehat{\kappa}_{\mathrm{C}} = (p_a - p_e)/(1 - p_e)$, where p_e is Cohen's percent chance agreement (see equation 3.3.1 of chapter 3).

- When there is no missing rating, and the ratings are organized in a contingency table then the variance of the unweighted Cohen's kappa coefficient is given by,

$$\begin{aligned} v(\widehat{\kappa}_{\mathrm{C}}) \quad &= \frac{1-f}{n(1-p_e)^2}\Bigg(p_a(1-p_a) - 4(1-\widehat{\kappa}_{\mathrm{C}})\sum_{k=1}^{q} p_{kk}\pi_k \\ &\quad + 4p_e\widehat{\kappa}_{\mathrm{C}}(1-\widehat{\kappa}_{\mathrm{C}}) + 4(1-\widehat{\kappa}_{\mathrm{C}})^2\sum_{k=1}^{q}\sum_{l=1}^{q} p_{kl}\pi_{kl}^2\Bigg), \end{aligned} \tag{6.3.7}$$

where $\pi_{kl} = (p_{+k} + p_{l+})/2$. A version of this expression was initially published by Gwet (2008a), and is mathematically equivalent to equation 13 of Fleiss et al. (1969), assuming no finite population correction (i.e. the sampling fraction f is 0).

- When there is no missing rating, and the ratings are organized in a contingency table, then the variance of the weighted Cohen's kappa coefficient is given by,

$$\begin{aligned} v(\widehat{\kappa}_{\mathrm{C}}) \quad &= \frac{1-f}{n(1-p_e)^2}\Bigg\{\sum_{k,l}^{q} p_{kl}\Big[w_{kl} - 2(1-\widehat{\kappa}_{\mathrm{C}})\overline{\pi}_{kl}\Big]^2 \\ &\quad - \Big[\widehat{\kappa}_{\mathrm{C}} - p'_e(1-\widehat{\kappa}_{\mathrm{C}})\Big]^2\Bigg\}, \end{aligned} \tag{6.3.8}$$

[14] There will be one column for each of the two raters, and a row represents the ratings associated with one subject.

where $\overline{\pi}_{kl} = (\overline{p}_{+k} + \overline{p}_{l+})/2$ (Note that the $\overline{\pi}_{kl}$ used here is different from that used to compute the variance of Scott's coefficient in equation 6.3.6). Note that the percent chance agreement p_e for the weighted Kappa coefficient is given by,

$$p_e = \sum_{k,l}^{q} w_{kl} p_{k+} p_{+l}.$$

- If your data contains missing ratings for some subjects, then I suggest to avoid summarizing it in a contingency table before analysis. Instead, the ratings from both raters must be listed explicitly and the more general variance equation of Conger's Kappa (c.f. equation 6.3.22) must be used.

Two-Rater Variances of the Unweighted and Weighted Brennan-Prediger Coefficient

The generalized G-index, also referred to as the Brennan-Prediger (BP) coefficient is given by $\widehat{\kappa}_{\text{BP}} = (p_a - 1/q)/(1 - 1/q)$.

- When there is no missing rating, and the ratings are organized in a contingency table then the variance of the *unweighted BP coefficient* is given by,

$$v\big(\widehat{\kappa}_{\text{BP}}\big) = \frac{1-f}{n(1-1/q)^2} p_a(1-p_a). \tag{6.3.9}$$

- When there is no missing rating, and the ratings are organized in a contingency table, then the variance of the *weighted BP coefficient* is given by,

$$v\big(\widehat{\kappa}_{\text{BP}}\big) = \frac{1-f}{n(1-p_e)^2} \sum_{k=1}^{q} \sum_{l=1}^{q} p_{kl}(w_{kl} - p_a)^2, \tag{6.3.10}$$

where p_a is the weighted percent agreement (i.e. the weighted sum of the p_{kl} values).

- If your data contains missing ratings then I recommend analyzing the raw ratings rather than the contingency table to avoid any loss of information. The more general variance equation 6.3.19 of the BP coefficient must be used.

Two-Rater Variance of the Unweighted and Weighted Krippendorff's $\widehat{\alpha}_{\text{K}}$

Krippendorff's alpha coefficient denoted by $\widehat{\alpha}_{\text{K}}$ is based solely on subjects that were rated by both raters. All subjects with a missing rating are downright excluded from analysis. For the purpose of calculating Krippendorff's alpha, n always

represents the count of subjects rated by both raters. This coefficient is given by $\widehat{\alpha}_{\text{K}} = (p_a - p_e)/(1 - p_e)$, where $p_a = (1 - \varepsilon_n)p'_a + \varepsilon_n$, $\varepsilon_n = 1/(2n)$, p'_a and p_e are given by,

$$p'_a = \sum_{k,l}^{q} w_{kl} p_{kl}, \text{ and } p_e = \sum_{k,l}^{q} w_{kl}\pi_k\pi_l. \tag{6.3.11}$$

$p_{kl} = n_{kl}/n$ is the relative number of subjects that raters A and B classified into categories k and l respectively. Moreover, $\pi_k = (p_{k+} + p_{+k})/2$ where p_{k+} = proportion of subjects that rater A classified into category k, and p_{+l} = proportion of subjects that rater B classified into category l.

The standard error of the weighted Krippendorff's alpha coefficient is obtained as the square root of its variance, which is defined as follows:

$$\begin{aligned} v\big(\widehat{\alpha}_{\text{K}}\big) \quad &= \frac{1-f}{n'(1-p_e)^2}\Bigg\{\sum_{k,l}^{q} p_{kl}\Big[w_{kl} - (1-\widehat{\alpha}'_{\text{K}})\overline{\pi}_{kl}\Big]^2 \\ &\quad - \Big[\widehat{\alpha}'_{\text{K}} - p_e(1-\widehat{\alpha}'_{\text{K}})\Big]^2\Bigg\}, \end{aligned} \tag{6.3.12}$$

where n' is the number of subjects rated by both raters, $\widehat{\alpha}'_{\text{K}} = (p'_a - p_e)/(1 - p_e)$, and $\overline{\pi}_k$, $\overline{\pi}_l$, and $\overline{\pi}_{kl}$ are defined as follows:

$$\overline{\pi}_k = \sum_{l=1}^{q} w_{kl}\pi_l, \overline{\pi}_l = \sum_{k=1}^{q} w_{kl}\pi_k, \text{ and } \overline{\pi}_{kl} = (\overline{\pi}_k + \overline{\pi}_l)/2. \tag{6.3.13}$$

Note that the quantity $\overline{\pi}_{kl}$ used here is different from that used to compute the variances of Scott and coefficient and that in equation 6.3.6).

This variance expression can be used for computing the weighted as well as the unweighted Krippendorff's coefficient. To get the unweighted coefficient, one needs to use identity weights. If your dataset contains missing ratings then the contingency table should not be used. Instead the raw ratings must be used along with the more general variance expression of Krippendorff's alpha of equation 6.3.21.

Two-Rater Variances of the Unweighted and Weighted Percent Agreement

The weighted percent agreement p_a is given by,

$$p_a = \sum_{k=1}^{q}\sum_{l=1}^{q} w_{kl} p_{kl}. \tag{6.3.14}$$

Used with identity weights, this expression will yield the regular unweighted percent agreement known in the literature.

- When there is no missing rating, and the ratings are organized in a contingency table, then the variance of the weighted percent agreement is given by,

$$v(p_a) = \frac{1-f}{n} \sum_{k=1}^{q} \sum_{l=1}^{q} p_{kl}(w_{kl} - p_a)^2. \tag{6.3.15}$$

- If your data contains missing ratings then rather than organizing it in the form of a summary contingency table, I recommend analyzing the raw ratings to avoid any loss of information. The more general variance expression of equation 6.3.20 is more appropriate in this situation.

Example 6.1

Let us once again consider the inter-rater reliability data of Table 3.4 in chapter 3. This data represents the distribution of patients with back pain classified into 3 pain categories by two clinicians 1 and 2. The inter-rater reliability experiment that produced it involves a sample of 102 participating patients randomly selected from a larger population of patients of interest suffering from back pain. However no clinician other than the two who participated in the experiment is of interest. Consequently, any extent of agreement obtained will only be applicable to the two participating clinicians. The standard error calculated in this situation must be conditional on this specific pair of participating clinicians. That is, it will measure the variation that is solely due to the selection of patients, the two patients being fixed.

Table 6.3: Percent of Patients (p_{kl}) with Back Pain by Pain Category and Clinician

Clinician 1	Clinician 2 Category				
Category	DER(%)	DYS(%)	POS(%)	$(p_{k+})^a$	$(\pi_k)^b$
DER	21.6	9.8	2.0	0.333	0.314
DYS	5.9	26.5	10.8	0.431	0.422
POS	2.0	4.9	16.7	0.235	0.265
$(p_{+k})^c$	0.294	0.412	0.294	1	1

[a] Clinician 2's category classification probability
[b] Average category classification probability
[c] Clinician 1's category classification probability

Table 6.4: Agreement Coefficients and Associated Standard Errors for Table 3.4 Data

Agreement Coefficients	Statistics			
	Coefficient	Standard Error	$(p_a)^a$	$(p_e)^b$
Gwet's AC_1	0.4757	0.070	0.6471	0.3269
Scott's Pi	0.4602	0.073	0.6471	0.3462
Cohen's Kappa	0.4613	0.073	0.6471	0.3449
Brennan-Prediger	0.4706	0.071	0.6471	0.2500
Krippendorff's Alpha	0.4628	0.073	0.6488	0.3462

[a] Percent Agreement
[b] Percent Chance Agreement

Table 6.3 contains various probabilities used in the standard error calculation, while Table 6.4 shows the magnitude of different agreement coefficients and their associated standard errors. The standard error is the most commonly-used precision measure in practice.

A standard error of 0.070 associated with an AC_1 coefficient of 0.4757 should be interpreted as follows: *"Our best estimate of the extent of agreement between clinicians 1 and 2 based on the AC_1 statistic and a sample of 102 patients is 0.4757. Its margin of error of 0.14 (i.e. 2×0.070) indicates that our best estimate may be off by 0.14, more or less."* In other words the "True" extent of agreement, which is based on all subjects in our target population could be as low as 0.3350 and as high as 0.6163. Some researchers may deem this range of values too wide, it reflects the limitations of the inter-rater reliability experiment that produced it. A remedy to this problem would be to increase the number of patients participating in the experiment. Guidelines for computing the optimal number of subjects are discussed in section 6.5.

Example 6.1 indicates that the margin of error associated with an agreement coefficient could be substantial. The following 3 important factors could contribute to its magnitude:

- The size n of the subject sample is a key contributing factor to the magnitude of the margin of error. An increase in number of participating subjects will lead to a decrease in the margin of error. However adding more subjects in an inter-rater reliability experiment will make it more expensive. Therefore a compromise must be found between the desired precision level for your estimate and the cost that you can afford for the experiment.

- A second important contributing factor is the size of our target population. Note that all variance expressions involve a multiplicative finite-population correction factor $1 - f = 1 - n/N$, which indicates that for a fixed sample size n, a smaller target population (i.e. smaller value for N) will lead to a smaller variance; and therefore to a smaller margin of error. Therefore for a limited budget, an inexpensive way to improve the accuracy of estimates is to limit the scope of the inter-rater reliability study by reducing the size of the target subject population.

- The last contributing factor is the magnitude of the "true" and unknown agreement coefficient. If the extent of agreement among raters is expected to be high, one may expect the margin of error to be small for the same sample size. This suggests that providing some training to the raters prior to the reliability experiment will improve the coefficient's precision in addition to increasing its magnitude.

Although very useful for computing the variances of inter-rater reliability coefficients when the number of raters is limited to 2, the variance equations presented in this section are not applicable to experiments involving 3 raters or more. This problem is addressed next.

6.3.1b) VARIANCES FOR MULTIPLE RATERS (2 OR MORE) BASED ON RAW RATINGS

Gwet (2008a) proposed variance estimators for the multiple-rater version of AC_1, and Fleiss' Kappa in addition to proving their validity with a Monte-Carlo simulation. The objective in this sub-section is to summarize these results, and to expand them in order to cover the weighted agreement coefficients, the missing ratings, as well as Conger's Kappa, and Krippendorff's alpha. In order to shorten the presentation of these results, we will show variance expressions only for the weighted versions of the coefficients. These expressions will cover unweighted coefficients as well by using identity weights.

Let n be the number of subjects rated by one rater or more, and n' the number of subjects rated by two raters or more. The percent agreement on a specific subject i is denoted by $p_{a|i}$ and is formulated as follows:

$$p_{a|i} = \sum_{k=1}^{q} \frac{r_{ik}(r_{ik}^{\star} - 1)}{r_i(r_i - 1)}, \text{ if } r_i \geq 2, \text{ and } p_{a|i} = 0 \text{ otherwise.} \tag{6.3.16}$$

You may want to refer to equation 6.2.4 for a definition of the different variables used in this equation.

▶ The AC_2 Coefficient

Let $f = n/N$ be the sampling fraction, and w_{kl} the set of weights to be used in the analysis. The AC_2 coefficient is defined by equation 5.6.1 of chapter 5, and its variance given by,

$$v\big(\widehat{\kappa}_{\mathrm{G}}\big) = \frac{1-f}{n}\frac{1}{n-1}\sum_{i=1}^{n}\big(\kappa^{\star}_{\mathrm{G}|i} - \widehat{\kappa}_{2\mathrm{G}}\big)^2, \tag{6.3.17}$$

where,

- $\kappa^{\star}_{\mathrm{G}|i} = \kappa_{\mathrm{G}|i} - 2(1-\widehat{\kappa}_{\mathrm{G}})\dfrac{p_{e|i}-p_e}{1-p_e}$,
- $\kappa_{\mathrm{G}|i} = \begin{cases} (n/n')(p_{a|i}-p_e)/(1-p_e), & \text{if } r_i \geq 2, \\ 0, & \text{otherwise}, \end{cases}$
- $p_{e|i} = \dfrac{T_w}{q(q-1)}\displaystyle\sum_{k=1}^{q}\frac{r_{ik}}{r_i}(1-\pi_k)$.

This expression may also be used to compute the variance of the AC_1 coefficient by using the appropriate set of weights (the identity weights).

▶ Fleiss' Kappa Coefficient $\widehat{\kappa}_{\mathrm{F}}$

The variance of Fleiss' Kappa coefficient (see section 4.4 of chapter 4 for a definition of the weighted Fleiss' coefficient) is given by,

$$v(\widehat{\kappa}_{\mathrm{F}}) = \frac{1-f}{n}\frac{1}{n-1}\sum_{i=1}^{n}\big(\kappa^{\star}_{\mathrm{F}|i} - \widehat{\kappa}_{\mathrm{F}}\big)^2, \tag{6.3.18}$$

where,

- $\kappa^{\star}_{\mathrm{F}|i} = \kappa_{\mathrm{F}|i} - 2(1-\kappa_{\mathrm{F}})\dfrac{p_{e|i}-p_e}{1-p_e}$,

  $\kappa_{\mathrm{F}|i} = \begin{cases} (n/n')(p_{a|i}-p_e)/(1-p_e), & \text{if } r_i \geq 2, \\ 0, & \text{otherwise}, \end{cases}$
- $p_{e|i} = \displaystyle\sum_{k=1}^{q}\overline{\pi}_k r_{ik}/r_i$, with $\overline{\pi}_k = (\overline{\pi}_{k+} + \overline{\pi}_{+k})/2$,

  $\overline{\pi}_{k+} = \displaystyle\sum_{l=1}^{q} w_{kl}\pi_l$, and $\overline{\pi}_{+l} = \displaystyle\sum_{k=1}^{q} w_{kl}\pi_k$.

The weight matrix is generally symmetric, in which case $\overline{\pi}_{k+} = \overline{\pi}_{+k}$.