CHAPTER 5

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) -

5.1 The Problem

Tables 5.1 and 5.2 are two representations of hypothetical rating data that Conger (1980) used as examples to illustrate the Kappa coefficient. The ratings are those of 4 raters $R1$, $R2$, $R3$, and $R4$ who each classified 10 subjects into one of 3 possible categories a, b, or c. Applied to this data, Fleiss' generalized Kappa (see equation 2.11 of chapter 2) yields an inter-rater reliability of $\widehat{\kappa}_{\text{F}} = 0.247$. To interpret the meaning of this empirical number, and to understand its real value, the researcher may need answers to some of the following fundamental questions:

- Is 0.247 a valid number ? Does it quantify the actual phenomenon the researcher wants to measure ? 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 ?"
- The Kappa coefficient of 0.247 is based on a single sample of 10 subjects and 4 raters. Are the 10 participating subjects sufficient in number to prove the reliability of a newly-developed classification system ? Assuming the number 0.247 measures what it is supposed to measure, how accurate is it ? Moreover, will the 4 raters of the study be the only ones to use the classification system ? How would a different group of raters affect inter-rater reliability ?

Asking those questions leads you straight to the domain 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 are based on the sampling distribution of the agreement coefficients of interest.

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 po-

pulation 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.

Table 5.1:
Categorization of 10 subjects into 3 groups $\{a, b, c\}$

Subjects	Raters			
	$R1$	$R2$	$R3$	$R4$
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 5.2:
Distribution of 4 Raters by Subject and Category

Subjects	Categories			Total
	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

The number of statistical techniques developed to address various practical problems is very large. Determining which ones apply to our particular problem, requires some efforts. The researcher must first and foremost develop a clear understanding of the reliability experiment's main objective. The following two objectives are often of interest:

- The researcher 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 score is seen as a sample[1] 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) from 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 an effective 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 of interest. These two target groups are generally bigger than what the researcher can afford to include in the 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[2] 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 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 the participating subjects were selected from. 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[3]", and will be the focus in this chapter.

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

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

[3] This framework for statistical inference was invented by a Polish mathematician named Jersey 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)

bounds are very conservative and should be expected to be at times substantially higher than the actual precision measures.

Table 5.10: Required Number of Raters by Desired Variation Coefficient

Desired Coefficient of Variation	Required Number Of Raters
5%	40
10%	20
15%	14
20%	10
25%	8
30%	7
40%	5
50%	4
70%	3
100%	2

5.6 Concluding Remarks

Chapter 5 presents a framework for statistical inference suitable for the analysis of inter-rater reliability data, and based on the randomization of the procedures for selecting subjects and raters. The main difference with alternative frameworks found in the inter-rater reliability literature is the absence of a theoretical statistical model. Statistical models commonly used in the literature are based on a hypothetical probability distribution associated with the ratings, and which serves as basis for calculating expectations and variances. A second difference appears in the way the randomization approach makes explicit the contribution of subjects and raters to the overall variation associated with the inter-rater reliability coefficients. Before the work of Gwet (2008*b*), the selection of raters was generally not treated as a source of variation affecting inter-rater reliability coefficients, except when the intraclass correlation coefficient was used for measuring the extent of agreement among raters.

In situations where the practitioner wants to extrapolate the findings to a larger universe of subjects, but not to any universe of raters larger than the group of participating raters, inference will be done conditionally on the rater sample made up of the specific raters who provided the ratings being analyzed. The validity of our analysis will then be limited to this particular group of raters. If the results of your analysis must be extrapolated to a larger universe of raters, but not to a universe of subjects larger than the group of participating subjects, inference is carried out conditionally upon the subject sample. Inference will be unconditional when the practitioner wants to extrapolate the findings to larger universes of raters

and subjects simultaneously. Consequently, three types of inferences (two conditional and one unconditional) can be implemented with any of the chance-corrected inter-rater reliability coefficients studied in the past few chapters. After providing a formal definition of the inter-rater reliability variance for each of the three types of inferences, we also proposed variance estimation procedures that practitioners can use with actual ratings. We saw through some examples, that raters alone may account for as much as 50% of the total variance, and should therefore not be neglected when computing the variance unless conditional inference is deemed appropriate.

The expressions used for computing the variance of the various agreement coefficients are often complex, and cannot be conveniently used to determine the optimal number of subjects or of raters when designing inter-rater reliability studies. In this chapter, I proposed simple rules of thumb that can be used to overcome these difficulties. These simple procedures will generally lead to sample sizes that produce adequate precision levels for all agreement coefficients, even though they were initially designed to control the variance of the percent agreement.