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 misreporting of ratings to deliberate misstatements by some raters, could affect the precision of Kappa or AC_1, 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 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.

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 2 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 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

[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.

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 2 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.

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

[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)

5.2 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, 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 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.

Defining 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.

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 the statistical inference.

In a target population of N subjects for example, the total number of samples of size n that one may form is the number of combinations[4] 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[5] of R raters in groups of r, equals the 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 a complex sampling design.

5.2.1 *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

[4]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 "$=\mathsf{COMBIN}(N, n)$"

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

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 in 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:

$$\pi_k = \frac{1}{N}\sum_{i=1}^{N} P_{ik}. \tag{5.1}$$

For the sake of clarity, let us consider Fleiss' generalized weighted Kappa as an example. The percent agreement probability and Fleiss' percent chance agreement (Fleiss, 1971), calculated at the population level 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 population, and not restricted to the samples), and defined as follows:

$$P_a = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{q}\frac{R_{ik}(R^{\star}_{ik}-1)}{R_i(R_i-1)}, \quad \text{and } P_e = \sum_{k,l} w_{kl}\pi_k\pi_l, \tag{5.2}$$

where $R^{\star}_{ik}$ is the weighted count of raters who classified subject i into any category linked to k through weighting (i.e. 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,

$$\kappa_{\text{F}} = \frac{P_a - P_e}{1 - P_e}. \tag{5.3}$$

All the quantities P_{ik}, π_k, P_e, P_a or γ_{π} are population parameters to be estimated from the subject and rater samples. We generally use capital Latin letters or Greek letters for population parameters, while sample-based estimated values of these parameters use small Latin letters, or capital Latin letters with a hat on the top. For example, $p_{ik} = r_{ik}/r_i$ is the estimated percent of raters who classified subject i into category k, with r_i being the number of sample raters who rated subject i. Similarly, the estimated values of P_a and P_e respectively denoted by p_a and p_e are defined as follows :

$$p_a = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{q}\frac{r_{ik}(r^{\star}_{ik}-1)}{r_i(r_i-1)}, \quad p_e = \sum_{k,l} w_{kl}\widehat{\pi}_k\widehat{\pi}_l, \tag{5.4}$$

where $r^{\star}_{ik}$ is the weighted count of sample raters who classified subject i into a category related to k through the weights, $\widehat{\pi}_k$ is 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 5.3 is denoted by $\widehat{\kappa}_{\text{F}}$ and given by:

$$\widehat{\kappa}_{\text{F}} = \frac{p_a - p_e}{1 - p_e}. \tag{5.5}$$

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 5.3)) 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 4.

For Gwet's AC_1 or AC_2, the associated population parameters will be respectively γ_1 and γ_2. Their sample-based estimates, respectively denoted by $\widehat{\gamma}_1$ and $\widehat{\gamma}_2$, are defined in chapter 4. Their variances are discussed in the next few sections.

5.2.2 *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 sampling error, which may be large. Consequently, some researchers will resort to interval estimation that provides a range of values (i.e. 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 the sample data. 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 statistics. 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 agreement value $p_a(s_n, s_r^\star)$. All $\mathrm{C}_{\mathrm{N}}^n \times \mathrm{C}_{\mathrm{R}}^r$ such pairs of samples lead to a series of $\mathrm{C}_{\mathrm{N}}^n \times \mathrm{C}_{\mathrm{R}}^r$ values $p_a(s_n, s_r^\star)$, which forms the sampling distribution of p_a upon which statistical inference will be 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.

5.3 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 5.3.1 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 agreement coefficient will solely be due to the sampling of subjects.

5.3.1 *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 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:

$$V\big(\widehat{\kappa}(s_n)|s_r^\star\big) = \sum_{b=1}^{\mathrm{C}_{\mathrm{N}}^{n}} P\Big(s_n^{(b)}\Big)\Big[\widehat{\kappa}\big(s_n^{(b)}\big) - \overline{\widehat{\kappa}}\Big]^2, \tag{5.6}$$

where $s_n^{(b)}$ is the b^{th} subject sample, $\overline{\widehat{\kappa}}$ is the average of all C_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 5.6 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 5.6 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 (5.6). 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.

5.3.1(a) Estimated Variances in Two-Rater Reliability Experiments

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 2.7 of chapter 2, 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.

Variances of the Unweighted and Weighted AC_1

Let $\widehat{\kappa}_G$ denote the AC_1 statistic[6]. It follows from chapter 4 that $\widehat{\kappa}_G = (p_a - p_e)/(1-p_e)$ 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)[7],

► When there is no missing rating, the variance of the unweighted AC_1 coefficient

[6] We will use $\widehat{\kappa}_G$ as the generic notation for Gwet's AC_1 and AC_2 coefficients. The unweighted $\widehat{\kappa}_G$ will refer to the AC_1, while the weighted $\widehat{\kappa}_G$ will refer to AC_2.

[7] 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.

proposed by Gwet (2008a) is given by:

$$v\big(\widehat{\kappa}_{\rm G}\big) = \frac{1-f}{n(1-p_e)^2}\Bigg\{p_a(1-p_a) - 4(1-\widehat{\kappa}_{\rm G})\Bigg(\frac{1}{q-1}\sum_{k=1}^{q}p_{kk}(1-\hat{\pi}_k) - p_a p_e\Bigg)$$
$$+ 4(1-\widehat{\kappa}_{\rm G})^2\Bigg(\frac{1}{(q-1)^2}\sum_{k=1}^{q}\sum_{l=1}^{q}p_{kl}\Big[1-(\pi_k+\pi_l)/2\Big]^2 - p_e^2\Bigg)\Bigg\}, \qquad (5.7)$$

▶ The following variance expression is more general and can be used to calculate the variance of the unweighted as well as the weighted AC_1 with the presence or absence of missing ratings,

$$v(\widehat{\kappa}_{\rm G}) = \frac{1-f}{n}\frac{1}{n-1}\sum_{i=1}^{n}(u_i-\overline{u})^2, \qquad (5.8)$$

where n is the number of subjects rated by at least one rater, and f the sampling fraction. Moreover, $u_i = a_i + 2(1-\widehat{\kappa}_{\rm G})e_i$, with a_i and e_i being defined as,

$$a_i = \frac{1}{(1-p_{\rm X})(1-p_e)}\sum_{k=1}^{q}\sum_{l=1}^{q}w_{kl}\Big[\delta_{kl}^{(i)} - (1-\delta_{\rm X}^{(i)})p'_{kl}\Big],$$
$$e_i = \frac{T_w}{q(q-1)(1-p_e)}\sum_{k=1}^{q}\pi_k b_k^{(i)},$$

where $\delta_{\rm X}^{(i)} = 1$ if only one of the 2 raters rated subject i, and $\delta_{\rm X}^{(i)} = 0$ otherwise (i.e both raters rated subject i). $b_k^{(i)} = \big(b_{+k}^{(i)} + b_{k+}^{(i)}\big)/2$ where,

$$b_{k+}^{(i)} = \frac{\delta_{k+}^{(i)} - (1-\delta_{\rm X+}^{(i)})p'_{k+}}{1-p_{\rm X+}}, \text{ and } b_{+k}^{(i)} = \frac{\delta_{+k}^{(i)} - (1-\delta_{+\rm X}^{(i)})p'_{+k}}{1-p_{+\rm X}}.$$

Note that $\delta_{\rm X+}^{(i)} = 1$ (resp. $\delta_{+\rm X}^{(i)} = 1$) if rater A (resp. rater B) did not score subject i and will be 0 otherwise.

Variances of the Unweighted and Weighted Scott's π Coefficient

Scott's π statistic (Scott, 1955) is given by $\widehat{\kappa}_{\rm S} = (p_a - p_e)/(1-p_e)$, where p_e is Scott's percent chance agreement of equation 2.6 in chapter 2.

▶ When there is no missing rating, the variance of the unweighted Scott's coefficient proposed by Gwet (2008a) is given by:

$$v\big(\widehat{\kappa}_{\rm S}\big) = \frac{1-f}{n(1-p_{e\pi})^2}\Bigg\{p_a(1-p_a) - 4(1-\widehat{\kappa}_{\rm S})\Bigg(\sum_{k=1}^{q}p_{kk}\pi_k - p_a p_e\Bigg)$$
$$+ 4(1-\widehat{\kappa}_{\rm S})^2\Bigg(\sum_{k=1}^{q}\sum_{l=1}^{q}p_{kl}\Big[(\pi_k+\pi_l)/2\Big]^2 - p_e^2\Bigg)\Bigg\}, \qquad (5.9)$$