CHAPTER 10

Inter-Rater Reliability: Conditional Analysis

OBJECTIVE

This chapter presents statistical techniques for analyzing the extent of agreement among raters conditionally upon the subject membership in a specific category. The specific category could be the subject's true category if it exists or the category into which one rater classified the subject. Conditional analysis offers the advantage of evaluating the extent of agreement among raters for a subgroup of subjects known to belong to a particular category. This analysis reduces the dependency of the agreement coefficient on trait prevalence and on the distribution of subjects across categories

CONTENTS

10.1 Overview **222**
10.2 Conditional Agreement Coefficient for 2 Raters in ACM Reliability Studies **224**
10.2.1 Basic Conditional Probabilities for ACM Studies **225**
10.2.2 Conditional Reliability Coefficient Between 2 Raters in ACM Reliability Studies **225**
10.2.3 Unconditional and Conditional Validity Coefficient Between 2 Raters in ACM Reliability Studies **227**
10.3 Conditional Agreement Coefficient for 3 Raters or More in ACM Studies **229**
▶ AC_1 as Validity Coefficient **230**
▶ Fleiss' Kappa as Validity Coefficient **231**
10.4 Conditional Agreement Coefficient for 2 Raters in RCM Studies **231**
10.4.1 Conditional Kappa Statistic of Light (1971) **234**
10.4.2 k-Conditional π-Statistic of Fleiss (1971) **236**
10.4.3 k-Conditional AC_1-Statistic **236**
10.4.4 k-Conditional Brennan-Prediger Statistic **237**
10.5 Concluding Remarks **238**

10.1 Overview

Scientific inquiries often involve classifying subjects into predefined categories. For example patients in a hospital could be labeled as "NORMAL" or "HIGH" according to their blood pressure level. In an inter-rater reliability experiment, category membership will be characterized either by a clear-cut operational definition establishing a deterministic relationship between subjects and categories, or by the raters' individual preferences. Clear operational definitions allow experts to determine the "true" score, also known in the literature as gold-standard scores, which are associated with each subject. The knowledge of true scores allow researchers to further investigate inter-rater reliability coefficients separately for each category, and to possibly identify problem categories where agreement is hard to reach. In this case, subjects are said to have an "Absolute Category Membership" (or ACM). When the categories are tied to the raters rather to the subjects, then classification depends more on each rater's preferences. No operational definition exists linking subjects to specific categories. The subjects are then said to have a "Relative Category Membership" (or RCM). Marginal probabilities in this case are often seen as fixed since raters have known preferences. Inter-rater reliability coefficients for RCM ratings could be further analyzed by considering only subjects that one rater classified into a specific category.

Let us consider an experiment involving chart review of women who show up in the Emergency Department with an abdominal pain or a vaginal bleeding. Two chart abstractors must assign 100 patients to one of the following two categories : (1) "Ectopic Pregnancy (EP)", and "Intrauterine Pregnancy (IUP)". However a highly experienced chart reviewer also categorizes all 100 patients into what is considered to be the "True"' categories. The results of this experiment are summarized in Table 10.1, where EP_T and IUP_T represent respectively the "True" (or Expert-ascertained) EP and IUP categories.

Table 10.1: Distribution of 100 Emergency Room Pregnant Women by Abstractor and Type of Pregnancy

Abstractor 1	Abstractor 2						Total		
	EP			IUP					
	EP_T	IUP_T	Total	EP_T	IUP_T	Total	EP_T	IUP_T	Total
EP	13	2	15	4	3	7	17	5	22
IUP	1	2	3	2	73	75	3	75	78
Total	14	4	18	6	76	82	20	80	100

Table 10.1 indicates that both abstractors categorized 15 pregnancies as Ectopic,

of which 13 are actually "True" Ectopic pregnancies while the other 2 are "True" Intrauterine pregnancies. Moreover, 14 of the 18 pregnancies that abstractor 2 classified as Ectopic are "True" Ectopic pregnancies while the remaining 4 are "True" IUPs.

It would be natural for a researcher to want to know whether abstractors would agree more easily when rating a "True" Ectopic pregnancy than when rating a "True" IUP. The conditional percent agreement given a True EP is $p_{a|\text{EP}} = (13+2)/20 = 0.75$ (i.e. the abstractors agreed to classify 13 of the 20 True EP as EPs, 2 as IUPs, and disagreed on the classification of the remaining 5 True EPs). Although the 2 true EPs classified as IUPs by both abstractors would increase reliability, they would certainly not increase validity, and should be excluded from consideration if validity is being measured. Validity will answer a research question such as "Would abstractors positively detect "True" Ectopic pregnancies more easily than they would detect true IUPs ? Identifying categories where agreement is more easily reached provides useful insight for further observer training, or for a possible modification of the response categories. This problem is resolved by breaking down the inter-rater reliability coefficient $\widehat{\kappa}$ into 2 components $\widehat{\kappa}_{\text{EP}}$, and $\widehat{\kappa}_{\text{IUP}}$ associated with the 2 response categories. These are two conditional inter-rater reliability coefficients further discussed in section 10.2.

Let us turn to reliability experiments where the notion of "True" scores is nonexistent. Consider Tables 10.2 and 10.3 where 2 raters classified 100 garments into one of 2 categories "Good" (or **G**) and "Bad" (or **B**). The rating process in this case depends more on the rater's personal taste than on the nature of the object. Even though the garment type still affects the rater's choice, the very relationship between the two remains under the rater's control. Consequently, rater's marginal probabilities can be considered fixed for a given population of garments, making them sufficiently important to play a pivotal role in the interpretation of the inter-rater reliability magnitude.

Table 10.2:
Distributions of 100 Garments by Rater (A/B) & Quality of Garment

Rater A's Scores	Rater B's Scores		
	B	**G**	**Total**
B	70	15	85
G	15	0	15
Total	85	15	100

Table 10.3:
Distributions of 100 Garments by Rater (C/D) & Quality of Garment

Rater C's Scores	Rater D's Scores		
	B	**G**	**Total**
B	50	40	90
G	0	10	10
Total	50	50	100

Based on the AC_1 coefficient, the extent of agreement between raters A and B

is evaluated at 0.597 and that between raters C and D evaluated at 0.31. Although AC_1 indicates that raters A and B are more in agreement than raters C and D by a ratio of almost 2 to 1, a close look at both Tables 10.2 and 10.3 suggests that given the observed marginal probabilities[1] raters A and B have reached the minimum agreement possible while raters C and D would have reached the maximum agreement possible. Therefore, one may argue that raters C and D are more in agreement than raters A and B (in a relative sense) given their respective rating propensities.

One objective of this chapter is to present ways to evaluate the extent of agreement among raters conditionally on their marginal probabilities. Conditional analysis of raters' agreement will generally be appropriate if the researcher wants to study the effect of categories on the agreement level, or if comparison between groups of raters is of interest and marginal probabilities can be assumed fixed.

10.2 Conditional Agreement Coefficient for 2 Raters in ACM Studies

Throughout this section, a k-subject refers to any subject whose "True" response category is k. The rating of subjects is said to be reliable when the raters consistently classify subjects into the same categories; but will be valid only if the subjects are consistently classified into their correct category by the raters. That is,

$$\mathsf{Validity} = \mathsf{Reliability} + \mathsf{Exactness}$$

In this section, we introduce reliability and validity measures. A measure of reliability in the case of 2 raters for example, represents the frequency with which both raters classify subjects into the same category (whether it is the "true" category or not). A measure of validity on the other hand shows how often both raters classify subjects into their "true" category. Because validity is a more stringent condition than reliability, validity coefficients are expected to be smaller than reliability coefficients. When the pool of subjects used to evaluate reliability or validity is restricted to k-subjects only, one obtains conditional reliability and conditional validity coefficients given category k. The use of all subjects for which ratings have been collected would lead to unconditional coefficients.

Throughout this chapter p_k refers to the probability that the "True" category of a randomly selected subject is k. Referring to the Emergency room pregnancy data of Table 10.1, the probability of "true" Ectopic pregnancy is $p_{\text{E}} = (14+6)/100 = 0.20$, while the probability of "true" Intrauterine Pregnancy is $p_{\text{I}} = (4+76)/100 = 0.80$

[1] i.e. the marginal probabilities (0.85 and 0.15 for rater A for instance) are considered fixed.

10.2.1 *Basic Conditional Probabilities for ACM Studies*

Conditional analysis of ACM data requires the use of various basic conditional probabilities that we will define in this section. To illustrate how they are calculated, we will occasionally use Table 10.1 data. These probabilities are defined as follows:

- $p_{ll}^{(k)}$: Probability for a randomly selected subject, to have k as its "True" category, and to be classified into the same category l by both raters. This is the "unconditional" agreement probability on category l for k-subjects.
- $p_{+l}^{(k)}$: Probability that a subject randomly selected from the subject universe turns out to be a k-subject and be classified into category l by rater 2.
- $p_{+l|k} = p_{+l}^{(k)}/p_k$: conditional probability for rater 2 to classify a subject into category l given that the subject's true category is k.
- The probabilities $p_{l+}^{(k)}$ and $p_{l+|k}$ could be defined in the same way as $p_{+l}^{(k)}$ and $p_{+l|k}$ have been, with the exception that this time both expressions would refer to rater 1 and not to rater 2. Table 10.1 data leads to the following probabilities:

$$\left(p_{+l}^{(k)}\right) = \left.\overbrace{\begin{pmatrix} 0.14 & 0.04 \\ 0.06 & 0.76 \end{pmatrix}}^{k=\text{E,I}}\right\}{\scriptstyle (+l)} \text{ and } \left(p_{l+}^{(k)}\right) = \left.\overbrace{\begin{pmatrix} 0.17 & 0.05 \\ 0.03 & 0.75 \end{pmatrix}}^{k=\text{E,I}}\right\}{\scriptstyle (l+)}, \quad (10.1)$$

- $\pi_{l|k}$: Probability for a randomly selected rater to classify into category l a subject randomly selected among k-subjects, and is calculated as follows :

$$\pi_{l|k} = \pi_{l(k)}/p_k \text{ where } \pi_{l(k)} = \left(p_{+l}^{(k)} + p_{l+}^{(k)}\right)/2. \quad (10.2)$$

Calculated from Table 10.1 data these probabilities take the following values :

$$\begin{pmatrix} \pi_{\text{E}(\text{E})} & \pi_{\text{E}(\text{I})} \\ \pi_{\text{I}(\text{E})} & \pi_{\text{I}(\text{I})} \end{pmatrix} = \begin{pmatrix} 0.155 & 0.045 \\ 0.045 & 0.755 \end{pmatrix},$$

and,

$$\begin{pmatrix} \pi_{\text{E}|\text{E}} & \pi_{\text{E}|\text{I}} \\ \pi_{\text{I}|\text{E}} & \pi_{\text{I}|\text{I}} \end{pmatrix} = \begin{pmatrix} 0.775 & 0.056 \\ 0.225 & 0.944 \end{pmatrix} \quad (10.3)$$

10.2.2 *Conditional Reliability Coefficient Between 2 Raters in ACM Reliability Studies*

Unconditional reliability coefficients for ACM studies are identical to the regular inter-rater reliability coefficients covered in part I of this book, and will not be further discussed in this section. This section will be devoted entirely to the

study of conditional reliability coefficients given the subject's "true" membership in a specific category. Considering Table 10.1, conditional reliability coefficients will for example quantify the extent of agreement among 2 raters under the condition that the women's "true" pregnancy type is Ectopic. As previously indicated, the primary purpose of conditioning is to have a reliability measure that is not affected by the distribution of women across the different types of pregnancy. Such a conditional reliability coefficient will facilitate comparison between reliability studies based on populations with different prevalence rates of Ectopic pregnancies for example.

► Let k be an arbitrary category (e.g. k could designate EP or IUP pregnancy). The Conditional AC_1 Reliability Coefficient given a "true" membership category k is denoted by $\widehat{\kappa}_{G|k}$, and defined as follows :

$$\widehat{\kappa}_{G|k} = \frac{p_{a|k} - p_{e|k}}{1 - p_{e|k}}, \text{ where } \begin{cases} p_{a|k} & = \dfrac{1}{p_k} \displaystyle\sum_{l=1}^{q} p_{ll}^{(k)}, \\ p_{e|k} & = \dfrac{1}{q-1} \displaystyle\sum_{l=1}^{q} \pi_{l|k}(1 - \pi_{l|k}). \end{cases} \tag{10.4}$$

Note that $p_{a|k}$ (the conditional percent agreement) is the probability that 2 raters agree given that the subject they both rated was randomly selected from the pool of k-subjects, while $p_{e|k}$ is the probability for both raters to agree by chance given a randomly selected k-subject.

► The Conditional Kappa Reliability Coefficient given a "true" membership category k is denoted by $\widehat{\kappa}_{C|k}$, and defined as follows :

$$\widehat{\kappa}_{C|k} = \frac{p_{a|k} - p_{e|k}}{1 - p_{e|k}}, \text{ where } p_{e|k} = \sum_{l=1}^{q} p_{+l|k} p_{l+|k}, \tag{10.5}$$

► The Conditional Pi reliability coefficient given a "true" membership category k is denoted by $\widehat{\kappa}_{S|k}$, and defined as follows :

$$\widehat{\kappa}_{S|k} = \frac{p_{a|k} - p_{e|k}}{1 - p_{e|k}}, \text{ where } p_{e|k} = \sum_{l=1}^{q} \pi_{l|k}^2, \tag{10.6}$$

► The Conditional Brennan-Prediger (BP) Reliability Coefficient given a "true" membership category k is denoted by $\widehat{\kappa}_{BP|k}$, and defined as follows :

$$\widehat{\kappa}_{BP|k} = \frac{p_{a|k} - p_{e|k}}{1 - p_{e|k}}, \text{ where } p_{e|k} = 1/q, \tag{10.7}$$