CHAPTER 11

Inter-Rater Reliability: Conditional Analysis

OBJECTIVE

This chapter introduces a number of measures of validity (as opposed to the measures of reliability discussed in the past few chapters), and describes statistical techniques for analyzing the extent of agreement among raters conditionally upon the subject membership in a specific category. This specific category used in the conditioning, 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, and can help identify a special group of subjects where agreement is hard to reach. Methods for computing the variances associated with these conditional measures are also discussed.

CONTENTS

11.1 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 than 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 generally 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 the chart review of women who enter the Emergency Department with an abdominal pain or a vaginal bleeding. Two chart abstractors named “Abstractor 1”, and “Abstractor 2” must assign 100 patients to one of the following two categories:

- Ectopic Pregnancy (EP), and
- Intrauterine Pregnancy (IP).

A highly experienced chart reviewer also categorizes the same 100 patients into what is considered to be the “True” categories. The results of this experiment are summarized in Table 11.1, where EP_{T} and IP_{T} represent respectively the “True” (or Expert-ascertained) EP and IP categories. Table 11.1 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” IPs.

It is natural for a researcher to want to know whether abstractors are more likely to agree while rating a “True” Ectopic pregnancy than while rating a “True” IP. Agreement in this case must be evaluated conditionally upon the true nature of the pregnancy. The statistically notion of conditioning applies in this case by restricting the pool of females subjects to be rated to those who carry a specific pregnancy type of interest. For example, the conditional percent agreement given a true EP is $p_{a|\mathrm{EP}} = (13+2)/20 = 0.75$. That is, abstractors agreed to classify 13 of the 20 true EPs as EPs, 2 as IPs, and disagreed about the classification of the remaining 5 True EPs. The denominator in this case is 20, because the analysis is limited to the 20 true EPs in the study group as shown in Table 11.1.

Table 11.1: Distribution of 100 Emergency Room Pregnant Women by Abstractor and Type of Pregnancy

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

Although the 2 true EPs classified as IPs by both abstractors would increase reliability, they would not increase validity. They should not be considered as agreement if validity is being measured. Validity will answer a research question such as "*Would abstractors more likely to positively detect true Ectopic pregnancies than they would positively detect true IPs ?*" Being able to identifying categories where agreement is more easily reached will identify other categories that should be the focus of further abstractor training. Conditional analysis could also lead to a possible modification of some categories that observers deem unclear. This analysis is carried out by breaking down the inter-rater reliability coefficient $\widehat{\kappa}$ into 2 components $\widehat{\kappa}_{\mathrm{EP}}$, and $\widehat{\kappa}_{\mathrm{IP}}$ associated with the 2 response categories. These two conditional inter-rater reliability coefficients are discussed in greater details in section 11.2.

Let us turn to reliability experiments where the notion of "True" scores is nonexistent. Consider Tables 11.2 and 11.3 where two raters classified 100 garments into one of two 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, the 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.

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 11.2 and 11.3 suggests that given the observed marginal probabilities[1], raters A and B have reached the minimum agreement possible while raters C and D 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.

Table 11.2:
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 11.3:
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

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.

11.2 Conditional Agreement Coefficient for two 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,

$$\textbf{Validity} = \textbf{Reliability} + \textbf{Exactness.}$$

In this section, I introduce reliability and validity measures. A measure of reliability in the case of two 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 quantifies the extent to which 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

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

$$\widehat{\kappa}_{\text{BP}|k} = \frac{p_{kk}/p_{2k} - p_{e|k}}{1 - p_{e|k}}, \text{ where } p_{e|k} = 1/q^2, \tag{11.4.16}$$

where q is the number of response categories.

Example 11.4

The conditional BP coefficients are illustrated in Table 11.18 using reliability data of Table 11.10. These coefficients consistently exceed 0.6. The "unconditional" BP-coefficient is given by $\widehat{\kappa}_{\text{BP}} = (0.917 - 0.5)/(1 - 0.5) = 0.834$.

Table 11.18: BP Agreement Coefficients Conditionally upon Category k by Judges A, B, and A or B Respectively

Category	Judge A			Judge B			Judges A or B		
(k)	$p_{a\|k+}$	$p_{e\|k+}$	$\widehat{\kappa}_{\text{BP}\|k+}$	$p_{a\|+k}$	$p_{e\|+k}$	$\widehat{\kappa}_{\text{BP}\|+k}$	$p_{a\|k}$	$p_{e\|k}$	$\widehat{\kappa}_{\text{BP}\|k}$
Bad	1.000	0.25	1.000	0.750	0.25	0.667	0.750	0.25	0.667
Good	0.889	0.25	0.852	1.000	0.25	1.000	0.889	0.25	0.852

11.5 Concluding Remarks

Although chance-corrected inter-rater reliability coefficients represented by a single index have been widely-accepted by researchers, they have also been criticized for being difficult to interpret. The difficulty in interpreting agreement coefficients stems from the dependency of inter-rater reliability coefficients on trait prevalence, or in general on the actual distribution of subjects by response category. It is the need to resolve this problem that led to the development of conditional agreement coefficients.

The notion of actual distribution of subjects by response category assumes for each subject the existence of a unique and specific category to which he can be classified objectively. I considered such subjects to have an Absolute Category Membership (ACM), and refer to them as ACM subjects. Reliability experiments involving ACM subjects were referred to as ACM studies, and agreement coefficients have been conditioned on the subjects' true category. I was able to quantify the extent of agreement among raters under the condition that the subjects belong to a certain category. These conditional agreement coefficients are expected to show more stability over time in addition to allowing comparison between different reliability studies. While rating an ACM subject, two raters may agree either about the correct category or about a wrong one. Although in both cases there will be agreement, I suggested that the extent of agreement about the true category be analyzed with validity coefficients and the extent of agreement about any category be analyzed with reliability coefficients. Both types of coefficients were also studied conditionally upon the subject's true membership category. Conditioning allows researchers to identify specific groups of subjects that prevent raters from reaching higher agreement levels.

As previously indicated, some reliability experiments are based on subjects that do not possess an Absolute Category Membership. For example experiments involving human subjects who express their preferences in the form of response categories. These subjects possess what I referred to as the Relative Category Membership (RCM). While agreement coefficients in ACM studies can be conditioned on the subject's true category, conditioning in RCM studies is typically done on the category into which one or more raters classified the subject. The AC_1 coefficient for example may quantify the extent of agreement conditionally upon subjects that rater 1 classified in category k. Such a coefficient in the context of two raters is denoted by $\widehat{\kappa}_{\mathrm{G}|k+}$. I have limited the conditional analysis in RCM studies to two raters only. This is due to the fact that the interest of practitioners for this type of analysis on RCM subjects is yet to be confirmed. Moreover, in a two-rater experiment the other rater is always special since it is necessarily the reference for comparison. In a multiple-rater experiment choosing one rater as reference for comparison might not always be justified. Although reliability coefficients are adjusted for chance agreement, conditioning works well if each rater can be assumed to have a reasonably stable rating pattern for a given subject population.

Practitioners would note that one limitation inherent to all conditional analyzes stems from the difficulty to have a precise evaluation of the various conditional probabilities. For ACM studies, a precise evaluation of the conditional probabilities would require the knowledge of each subject's true membership category in the subject universe. Although experts may at times provide that information, in most practical applications that information will not be available. In RCM studies, a precise evaluation of the conditional probabilities requires the knowledge of raters' marginal probabilities. That is the category into which each rater would classify each population subject. Because of the limited information available following a reliability experiment, the conditional probabilities are generally estimated using sample information with the risk of increasing the sampling error associated with the agreement coefficients.