CHAPTER 7

Intraclass Correlations in One-Factor Studies

OBJECTIVE

The objective in this chapter is to present methods for calculating the intraclass correlation coefficient and associated precision measures for single-factor reliability studies. We will consider situations where the quantitative outcome is studied as a function of either the rater effect or the subject, but not both. Intraclass correlation is first defined as an abstract construct before describing the computation procedures. Methods for obtaining confidence intervals and p-values will be presented as well.

CONTENTS

7.1 What is the Issue ?

In the past few chapters of part I, we presented many techniques for quantifying the extent of agreement among raters. Although some of these techniques were extended to interval data, the primary focus has been on nominal and ordinal data. This chapter as well as the other chapters of part II, are devoted to the study of inter-rater reliability for quantitative outcomes. Studying inter-rater reliability for quantitative outcomes amounts to studying their reproducibility.

Why do we need to care about intraclass correlation when weighted versions of the chance-corrected measures can be used to handle quantitative outcomes ? It is because the notion of "perfect" agreement that characterizes 2 raters assigning the exact same score to the same subject, does not translate well to quantitative measurements. Consider 2 electronic devices used to measure the knee joint laxity on 15 subjects. Even if both devices are equally reliable, we would not expect them to produce the exact same quantitative measurement on the same subjects, since these values belong to a continuum. Likewise, 2 raters that measure the height or the weight of a human subject will likely produce slightly different numbers regardless of their proficiency in the use of the measuring instrument. With agreement no longer referring to an exact match, the notions of chance agreement and percent agreement evaporate.

The solution to this problem is to use the portion of variation in the data that is due to subjects, and to compare it to the other portion of that variation due to raters. If the rater-induced variation exceeds that of the subject by a wide margin, the raters are said to have a low inter-rater reliability. Otherwise, the raters are said to have high inter-rater reliability. But this approach will work only if the reliability experiment is designed in such a way that the different variation components can be separated. We will see in the next few sections how this task can be accomplished. Several approaches can be used to design an inter-rater reliability study, depending on the goal aimed at for the study. In the next section, we will describe a few designs commonly used in the context of inter-rater reliability analysis.

7.2 The Design of Reliability Studies

Consider the reliability data shown in Table 7.1. That data represents scores that 4 raters assigned to 6 subjects, and could be interpreted in various ways depending on how it was collected. Here are 4 possible study designs (or data models) that could have produced Table 1 data:

- **Model 1A**: *Each subject is rated by a different group of raters*
  According to this model, each row of Table 7.1 is not necessarily associated with

the same set of 4 raters. Although the 4 raters are consistently labeled as raters 1, 2, 3, and 4, they could represent different individuals, or different measuring instruments. One may average the data row-wise to study the subject effect, but will not be able to average column-wise to obtain the rater effect. This is why this is often known as a one-factor (or one-way) model.

The main implication of this model is that one rater may not have the opportunity to score more than one subject. Consequently, this model makes it impossible to evaluate *Intra-rater Reliability*, which is a measure of the rater's self-consistency. However, the raters under this model still score the same subjects, making it possible to compute *Inter-rater reliability.*

Table 7.1: Scores assigned by 4 raters to 6 subjects[a]

	Rater				
Subject	1	2	3	4	Average
1	9	2	5	8	6
2	6	1	3	2	3
3	8	4	6	8	6.5
4	7	1	2	6	4
5	10	5	6	9	7.5
6	6	2	4	7	4.75
Average	7.67	2.5	4.33	6.67	5.29

[a]This data is taken from Shrout & Fleiss (1979), although we have replaced the terms Target and Judge with Subject and Rater respectively, and added row and column marginal averages.

- **Model 1B**: *Each rater scores a different group of subjects*
  If Table 7.1 data were collected according to this design, then the 6 subjects may differ from rater to rater. That is, each rater scored his own set of subjects, even though we may have decided to consistently labeled them as 1, 2, 3, 4, 5, and 6. One may control for the rater effect by averaging Table 7.1's columns. Any row-wise averaging would be meaningless as such an operation would involve different subjects as well as different raters. Therefore, the only factor that can be studied is the rater factor, and this model will later be referred to as a one-factor or one-way model.

  The main implication of this model is that it allows for the evaluation of intra-rater reliability, and not that of inter-rater reliability. Evaluating inter-rater reliability always require different raters to score the same subjects .

- **Model 2**: *The Random Factorial Design*
  According to this model, each subject is scored by the same group of raters. Both the subjects and the raters are random samples selected from the respec-

tive populations they represent, hence the naming "random" design. Moreover, the column and row marginal averages are meaningful, and the effects of subject and rater factors can be evaluated. It is because both factors (rater and subject) can be studied that this design is known as a "factorial design".

- **Model 3**: *The Mixed Factorial Design*
  According to this design, each subject is scored by the same group of raters, and is also in this regard a factorial design. Unlike Model 2, here only the group of subjects represents a random sample selected from a larger subject population, while the group of raters does not represent a random sample. Because the set of raters that participated in the reliability experiment was not randomly selected from a larger rater population, these raters only represent themselves. The resulting inter-rater reliability coefficient can therefore not be applied to raters beyond those in the experiment.

Each of these models requires a different method for calculating the intraclass correlation coefficient. Shrout & Fleiss (1979) discussed models 1A (although it was referred to as model 1), 2, and 3. The same models were also discussed by McGraw and Wong (1996), who presented methods for computing the intraclass correlation for each of them.

This chapter is devoted entirely to the study of models 1A and 1B. Models 2 and 3 will be discussed separately in subsequent chapters.

7.3 Intraclass Correlation under Model 1A

Let us consider a reliability experiment where r raters must each take m measurements (or replicates) on n subjects. One could say with respect to Table 7.1 that $n = 6$, $r = 4$, and $m = 1$, since there are 6 subjects, 4 raters, and 1 replicate (i.e. there is a single measurement taken by the raters on each subject. Let y_{ijk} be the abstract representation of the quantitative score assigned to subject i by rater j on the k^{th} occasion. The rater may change from subject to subject as stipulated in model 1A. The mathematical translation of this model is as follows:

$$y_{ijk} = \mu + s_i + e_{ijk}, \tag{7.1}$$

where μ is the overall mean score, s_i is subject i's effect, and e_{ijk} the error effect. Both effects are assumed to be random, independent[1] and to follow the Normal distribution with mean 0, and variances σ_s^2, and σ_e^2 respectively.

[1]Independence is taken here in a statistical sense. That is the knowledge of the magnitude of one effect tells nothing about the magnitude of the other effect

7.3.1 *Defining Inter-Rater Reliability for Individual Ratings*

The *Intraclass Correlation Coefficient* (ICC) needed to measure inter-rater reliability is by definition the correlation coefficient between the 2 quantitative scores y_{ijk} and $y_{ij'k}$ associated with the same subject i, and the same replicate number k, but with 2 raters j and j'. It follows from equation 7.1 that this particular correlation coefficient (denoted by ρ) is given by,.

$$\rho = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_e^2} \tag{7.2}$$

Equation 7.2 provides the theoretical definition of ICC as the ratio of the subject variance to the total variance (i.e. the sum of subject and error variances) based on model 7.1. This ratio shows that the ICC will be high when the subject variance exceeds the error variance by a wide margin. This quantity indeed represents the extent of agreement among the r raters. To see this, you must realize that the error variance σ_e^2 is actually the variance of 2 factors blended together, which are the rater factor and the error factor. However, the design represented by model 1A makes it impossible to separate them[2]. Therefore, a small error variance under model 1A, actually means that both the error and rater variances have to be small. It is that small (unknown) rater variance that ensures a small variation in the rater's scores and a high inter-rater reliability.

7.3.2 *Calculating Inter-Rater Reliability*

Shrout and Fleiss (1979) as well as McGraw and Wong (1996) presented ways to actually compute the intraclass correlation from raw data. Their methods are based on the use on various means of squares, and do not account for the presence of missing values. This could be problematic in practice as missing values are common in many applications. However, the use of the means of squares is particularly useful for planning purposes, and will help determine the required sample sizes, and number of replicates prior to the actual conduct of the study. Consequently, in this section we will present the computation methods needed to analyze data that have already been collected, and that may contain missing values (this is unbalanced data).

The approach that we present here is a simplification of the methods used by Searle (1997, page 474). Let m_{ij} be the number of measurements (or replicates) associated with subject i and rater j. In the case of Table 7.1, $m_{ij} = 1$ for all

[2]You would separate the rater and error variances only if each rater scores a whole set of subjects, which is not the case under model 1A

subjects and all raters. If rater j does not score subject i then $m_{ij} = 0$, indicating that these particular ratings are missing. Let M be the total number measurements collected for the whole study (i.e. M is the summation of all the m_{ij}'s). For Table 7.1, $M = 6 \times 4 = 24$. Here are a few quantities that we are going to need:

- $m_{i\cdot}$ = number of measurements associated with subject i. In Table 7.1 there are 4 values associated with each subject since none is missing. That is $m_{1\cdot} = m_{2\cdot} = \cdots = m_{6\cdot} = 4$.
- $m_{\cdot j}$ = number of measurements associated with rater j. In Table 7.1, there are 6 values associated with each rater since none is missing. That is, $m_{\cdot 1} = m_{\cdot 2} = m_{\cdot 3} = m_{\cdot 4} = 6$.
- $y_{i\cdot\cdot}^2$ is the square of the sum of all values associated with subject i, and T_{2s}, the summation of all $y_{i\cdot\cdot}^2$ values.
- Let T_y be the summation of all y_{ijk} scores, and T_{2y} the summation of all squared scores y_{ijk}^2.

The ICC of equation 7.2 is obtained by calculating the 2 variance components from the raw experimental data. While the theoretical subject variance is σ_s^2, its calculated value is denoted by $\widehat{\sigma}_s^2$ (read sigma hat s square). Likewise, the calculated error variance is denoted by $\widehat{\sigma}_e^2$. The calculated intraclass correlation coefficient is given by,

$$\boxed{\mathsf{ICC}(1\mathsf{A},1) = \frac{\widehat{\sigma}_s^2}{\widehat{\sigma}_s^2 + \widehat{\sigma}_e^2},} \tag{7.3}$$

where,

$$\widehat{\sigma}_e^2 = (T_{2y} - T_{2s})/(M - n), \tag{7.4}$$

$$\widehat{\sigma}_s^2 = \frac{T_{2s} - T_y^2/M - (n-1)\widehat{\sigma}_e^2}{M - k_0}, \tag{7.5}$$

and k_0 is calculated by summing all factors $m_{ij}^2/m_{\cdot j}$ over all n subjects and all raters.

Example 7.1

To illustrate the calculation of ICC(1A,1)[3], let us consider the data of Table 7.1 and assume that it was collected following the 1A design. Tables 7.2 and 7.3 show the different steps for calculating the intraclass correlation coefficients. Table 7.3 aims at showing the calculation of $k_0 = 4$, obtained by summing the last column. The last row

[3] "1A" in this notation indicates that ICC is based on model 1A, and the number 1 on the right side of the comma sign indicates that each rating used in the analysis represents 1 measurement, as opposed to being an average of several measurements.