CHAPTER 4

Intraclass Correlations in One-Factor Studies

OBJECTIVE

This chapter presents methods and techniques for calculating the intraclass correlation coefficient and associated precision measures in single-factor reliability studies based on models 1A and 1B. I consider situations where the quantitative measurement is studied as a function of either the rater effect or the subject effect, but not both. Intraclass correlation is first defined as an abstract construct before the computation procedures are described. Methods for obtaining confidence intervals and p-values will be presented as well. I also discuss some approaches for calculating the optimal number of subjects and raters needed while planning an inter-rater reliability experiment.

Contents

4.1 Intraclass Correlation under Model 1A

Let us consider a reliability experiment where r raters must each take m measurements (or replicate measurements or trial measurements) from n subjects. However, each subject could be rated by a different group of r raters. One could say with respect to Table 4.1 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} trial. The rater may change from subject to subject as stipulated in model 1A. The mathematical translation of this model is as follows:

$$y_{ijk} = \mu + s_i + e_{ijk}, \tag{4.1.1}$$

where μ is the expected 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.

4.1.1 *Defining Inter-Rater Reliability*

The *Intraclass Correlation Coefficient* (ICC) needed to measure inter-rater reliability is by definition the correlation coefficient between the two quantitative scores y_{ijk} and $y_{ij'k}$ associated with the same subject i, and the same replicate number k, but with two raters j and j'. It follows from equation 4.1.1 that this particular correlation coefficient (denoted by ρ) is given by,

$$\boxed{\rho = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_e^2}.} \tag{4.1.2}$$

Equation 4.1.2 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 4.1.1. 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 first realize that the error variance σ_e^2 is actually the variance of two factors blended together, which are the rater factor and the error factor. However, the design represented by model 1A

[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. Note that the effects s_i and $s_{i'}$ associated with 2 distinct subjects i and i' are independent as well. Likewise, all error effects e_{ijk} are assumed to be pairwise independent.

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 (and unknown) rater variance that ensures a small variation in the rater's scores and a high inter-rater reliability.

4.1.2 *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 assume that your dataset is complete (i.e. does not contain 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 conducting the actual study (see section 4.4). This section focuses on computation methods needed to analyze data already collected, and that may contain missing values.

The approach that I present here is a simplified version of the methods described 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 4.1, $m_{ij} = 1$ for all subjects and all raters. If rater j does not score subject i then $m_{ij} = 0$, indicating that these ratings are missing. Let M be the total number measurements collected for the whole study (i.e. M is the summation of all m_{ij} values). For Table 4.1, $M = 6 \times 4 = 24$. Here are a few quantities that we are going to need:

- $m_{i.}$ = number of measurements associated with subject i. In Table 4.1 there are 4 values associated with each subject since none is missing. That is $m_{1.} = m_{2.} = \cdots = m_{6.} = 4$.
- $m_{.j}$ = number of measurements associated with rater j. In Table 4.1, there are 6 values associated with each rater since none is missing. That is, $m_{.1} = m_{.2} = m_{.3} = m_{.4} = 6$.
- $y_{i.}^2$ is the squared value of subject i's total score, and T_{2s} the summation of all ratios $y_{i.}^2/m_{i.}$ over all subjects $i = 1, \cdots, n$.
- Let T_y be the total score (i.e. the summation of all y_{ijk} values), and T_{2y} the total sum of squares (i.e. the summation of all squared scores y_{ijk}^2).

In practice, the ICC of equation 4.1.2 can only be approximated using experimental data. This is done by calculating the two variance components from the raw

[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

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 associated with model 1A is denoted by ICC(1A,1)[3]

$$\boxed{\text{ICC}(1\text{A},1) = \frac{\widehat{\sigma}_s^2}{\widehat{\sigma}_s^2 + \widehat{\sigma}_e^2},} \tag{4.1.3}$$

where,

$$\widehat{\sigma}_e^2 = (T_{2y} - T_{2s})/(M - n), \tag{4.1.4}$$

$$\widehat{\sigma}_s^2 = \frac{T_{2s} - T_y^2/M - (n-1)\widehat{\sigma}_e^2}{M - k_0}, \tag{4.1.5}$$

and k_0 is the expected number of measurements per subject and is calculated by summing all factors $m_{ij}^2/m_{.j}$ over all n subjects and all raters. If there is no missing rating then $k_0 = rm$.

Example 4.1

To illustrate the calculation of ICC(1A,1), let us consider the data of Table 4.1 and assume that it was collected following the 1A design. Tables 4.1 and 4.2 show the different steps for calculating the intraclass correlation coefficients. Table 4.2 aims at showing the calculation of $k_0 = 4$, obtained by summing the last column. Columns 6, 7, 9, and 10 of Table 4.1 show the numbers $T_y = 127$, $M = 24$, $T_{2s} = 728.25$, and $T_{2y} = 841$. It follows from equations 4.1.4 and 4.1.5 that,

$$\widehat{\sigma}_e^2 = (841 - 728.25)/(24 - 6) = 6.264,$$
$$\widehat{\sigma}_s^2 = (728.25 - 127^2/24 - (6-1) \times 6.264)/(24 - 4) = 1.2444.$$

After plugging these variance components into equation 4.1.3, one obtains the intraclass correlation, $\text{ICC}(1A, 1) = 1.2444/(1.2444 + 6.264) = 0.1657$.

Interested readers may download the Excel spreadsheet,
www.agreestat.com/books/icc5/chapter4/chapter4examples.xlsx

which shows the step-by-step calculation of this ICC.

[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. Some authors (e.g. Shrout and Fleiss (1979) discussed the situation where the rating being analyzed is an average of k measurements. Although the magnitude of the ICC based on mean scores is higher, so is its standard error. It is unclear whether a higher and less accurate ICC should be favoured to a lower and more accurate ICC.

The approach used in Example 4.1 for computing ICC$(1A, 1)$ yields the exact same answer as the standard approach based on means of squares advocated by Shrout and Fleiss (1979), or McGraw and Wong (1996). Their standard approach however, can only work when there is no missing rating, and is given by,

$$\text{ICC}(1A, 1) = \frac{\text{MSS} - \text{MSE}}{\text{MSS} + (rm - 1)\text{MSE}}, \tag{4.1.6}$$

where MSS is the mean of squares for subjects, and MSE the mean of squares for errors. These means of squares are calculated as follows:

- MSS is calculated by summing all squared differences $(\overline{y}_{i\cdot\cdot} - \overline{y})^2$ over all n subjects, and by multiplying the summation by $rm/(n-1)$. Note that $\overline{y}_{i\cdot\cdot}$ represents the average of all measurements associated with subject i, while $\overline{y}$ is the overall average.

$$\text{MSS} = \frac{rm}{n-1}\sum_{i=1}^{n}(\overline{y}_{i\cdot\cdot} - \overline{y})^2 \tag{4.1.7}$$

- MSE is calculated by summing all squared differences $(y_{ijk} - \overline{y}_{i\cdot\cdot})^2$ over all n subjects, and by dividing the summation by $n(rm-1)$.

$$\text{MSE} = \frac{1}{n(rm-1)}\sum_{i=1}^{n}\sum_{j=1}^{r}\sum_{k=1}^{m}(y_{ijk} - \overline{y}_{i\cdot\cdot})^2 \tag{4.1.8}$$

Note that equations 4.1.3 and 4.1.6 are equivalent when there is no missing rating, in which case the variance components can be written as follows:

$$\widehat{\sigma}_e^2 = \text{MSE}, \text{ and } \widehat{\sigma}_s^2 = \frac{\text{MSS} - \text{MSE}}{rm}. \tag{4.1.9}$$

While the ICC calculation using a balanced dataset[4] of ratings rests entirely on one method described by equation 4.1.6, there are several valid methods for performing the same calculation with unbalanced datasets of ratings. Equation 4.1.3 is one (i.e. likely the simplest) of several methods for calculating the ICC for unbalanced data. The variance components in equation 4.1.3 are calculated using the Analysis of Variance method also known in the statistical literature as Henderson's Method 1. Researchers using other statistical packages need to request the Henderson's method 1 in order to replicate the calculations shown in Example 4.1.

[4] A balanced dataset of ratings is a complete dataset with no missing rating, while incomplete datasets are said to be unbalanced.

Table 4.1: Ratings of 6 Subjects from 4 Raters, and Calculation of ICC(1A,1) based on equation 4.1.3

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Subject (i)	Rater (j)								
	1	2	3	4	$y_{i\cdot\cdot}$	$m_{i\cdot}$	$y_{i\cdot\cdot}^2$	$y_{i\cdot\cdot}^2/m_{i\cdot}$	$y_{2i\cdot\cdot}$ [a]
1	9	2	5	8	24	4	576	144	174
2	6	1	3	2	12	4	144	36	50
3	8	4	6	8	26	4	676	169	180
4	7	1	2	6	16	4	256	64	90
5	10	5	6	9	30	4	900	225	242
6	6	2	4	7	19	4	361	90.25	105
Rating Sum/Rater $y_{\cdot j\cdot}$	46	15	26	40	T_y 127	M 24		T_{2s} 728.25	T_{2y} 841
# Ratings/Rater $m_{\cdot j}$	6	6	6	6	*Column Totals*				

[a]Note that $y_{2i\cdot\cdot}$ in the last column represents the row-wise summation of the squared raw values.

Table 4.2: Calculating k_0 using the values $m_{ij}^2/m_{\cdot j}$

Subject (i)	Rater (j)				
	1	2	3	4	Total
1	0.16667	0.16667	0.16667	0.16667	0.66667
2	0.16667	0.16667	0.16667	0.16667	0.66667
3	0.16667	0.16667	0.16667	0.16667	0.66667
4	0.16667	0.16667	0.16667	0.16667	0.66667
5	0.16667	0.16667	0.16667	0.16667	0.66667
6	0.16667	0.16667	0.16667	0.16667	0.66667
					$k_0 = 4$

4.1.3 *Defining Intra-Rater Reliability*

A question one may ask is whether an intraclass correlation coefficient defined under model 1A can adequately measure intra-rater reliability[5]. In general, the answer would be no. However, there is an exception that would allow experimental

[5]Intra-rater reliability represents the extent of self-consistency of the raters.

data generated under Model 1A to be used for calculating intra-rater reliability. This exception is an experiment with replication, involving two trials or more that are carried out on each subject. Such an experiment allows equation 4.1.2 to be interpreted as an intra-rater reliability coefficient as well. The intraclass correlation coefficient is then defined as the correlation coefficient between two replicate measurements y_{ijk} and $y_{ijk'}$. This coefficient is also known as the test-retest reliability coefficient.

Under model 1A, the notions of rater and replicate are confounded. That is, two ratings associated with the same subject are treated in the same way whether they were generated by two raters or by the same rater on two occasions. This explains why the same equation 4.1.2 represents both the inter-rater and the intra-rater reliability coefficient. Although this general reliability measure can serve several purposes, it also possesses several drawbacks. For example, if equation 4.1.2 is used as an intra-rater reliability coefficient and yields a low value, then you may never know whether this is due to a low reproducibility or whether it is due to a low agreement among raters. To resolve this problem, some researchers design intra-rater reliability experiments with a single rater and a number of replicate measurements per subject. This approach will work fine except that it is based on a single rater, which means that the use of a different rater will potentially change the intra-rater coefficient substantially. This will not be a problem if the different raters are already known to have high inter-rater reliability.

4.1.4 *Recommendations*

The use model 1A is highly recommended if the reliability experiment is designed in such a way that each subject is rated by multiple raters, and the researcher cannot guarantee that the same group of raters will be available to rate more than one subject. If that is the case, I recommend that only inter-rater reliability be calculated. Equation 4.1.2 is not a good measure of intra-rater reliability.

If the main study objective is to assess intra-rater reliability, then I would recommend using model 1B of section 4.2. Model 1B requires the use of two raters or more, and yields a valid intra-rater coefficient.

4.2 Intraclass Correlation under Model 1B

Under model 1B, different raters may rate different groups of subjects. Although there could be some overlap between the different groups of subjects, this is not a requirement under model 1B. This model is different from model 1A where each subject is generally rated by 2 raters or more, although the raters may differ from subject to subject. Under model 1A, you start with a group of subjects and randomly assign groups of raters of a given size with no possibility of knowing at