CHAPTER 5

Intraclass Correlations under the Random Factorial Design

OBJECTIVE

This chapter aims at presenting methods for calculating various intraclass correlation coefficients and associated precision measures when the rater and subject factors are fully crossed. Each rater is expected to rate all subjects, but may take more measurements on some subjects and less on others. The rater and subject samples are both assumed to have been randomly selected from larger rater and subject populations. I define two types of intraclass correlation coefficients(ICC): (i) the ICC for quantifying inter-rater reliability, and (ii) the ICC for quantifying intra-rater reliability. For both types of intraclass correlation coefficients, methods for obtaining confidence intervals, p-values, and optimal sample sizes (i.e. required number of subjects and raters during the design of experiments) are presented as well.

Contents

5.1 The Issues

The Intraclass Correlation Coefficient (ICC) associated with Model 1A, is the ratio of the subject variance to the sum of the subject and error variances. What was termed error variance in the previous chapter is in reality the variance of a combination of three effects, which are the rater effect, a possible rater-subject interaction effect[1] and the experimental error effect. Because these three effects are blended together, they are interdependent, and their combined variance is expected to be higher than if the experiment was designed to keep them independent[2]. Therefore, the researcher can improve the magnitude of the ICC substantially by designing the experiment so as to keep all the factors at play independent from one another. This is accomplished by getting each rater to score all subjects. Such a design is known as the factorial design and is the subject of this chapter.

The ICC associated with Model 1B on the other hand, quantifies the intra-rater reliability and was defined in the previous chapter as the ratio of the rater variance to the sum of the rater and error variances. Once again, the error variance in the context of model 1B is actually the variance of the combined effect due to the subject, the rater-subject interaction and the experimental error. The experimental design that underlies model 1B (i.e. each rater scores a different group of subjects) has blended these three effects into one. Consequently, the variance of the combined effect will often be high, reducing thereby the magnitude of the ICC. If an experiment is designed so that the rater, rater-subject interaction, and error effects are independent from one another, then the variance due to their interdependency will be eliminated leading to a higher ICC for the same amount of data collected. This is the factorial design mentioned in the previous paragraph.

There are different types of factorial designs that may achieve different objectives. We will now review some of them.

Types of Factorial Designs

The factorial design is an experimental design where each rater is expected to rate all subjects participating in the experiment. The main advantage of this design is that all the factors involved in the experiment are kept independent from one another. That is, you can fix a specific rater and study the subject effect; just as you may fix a specific subject so as to study the rater effect. If two measurements or more

[1] The rater-subject interaction can be seen as the portion of the rater effect that may be attributed to the specific subject being rated.

[2] Note that if a and b are 2 dependent effects, then their combined variance will be $var(a+b) = var(a) + var(b) + 2cov(a,b)$, where $cov(a,b)$ is the covariance between a and b. If the effects are independent, the covariance term will vanish, the joint variance will decrease (assuming a positive covariance, which is usually the case in agreement studies).

are taken from one subject by the same rater, then one may study the rater-subject interaction effect independently from the experimental error.

Rater-subject interaction is bad for both inter-rater and intra-rater reliability, but is sometimes unavoidable. It induces more variation in the data, in addition to the portion of total variation that is due to raters and subjects. This extra variation will further reduce the magnitude of the ICC. Figure 5.1 depicts the reliability data of Table 4.1 of chapter 4. Without interaction, all 4 curves associated with the raters would be reasonably parallel, which is the case for raters 1, 2, and 3. Rater 4 however, appears to assign scores to subjects with a gap with other raters that changes from subject to subject. This is an indication of the existence of rater-subject interaction. Rater 4 alone is likely to bring the ICC down in a significant way.

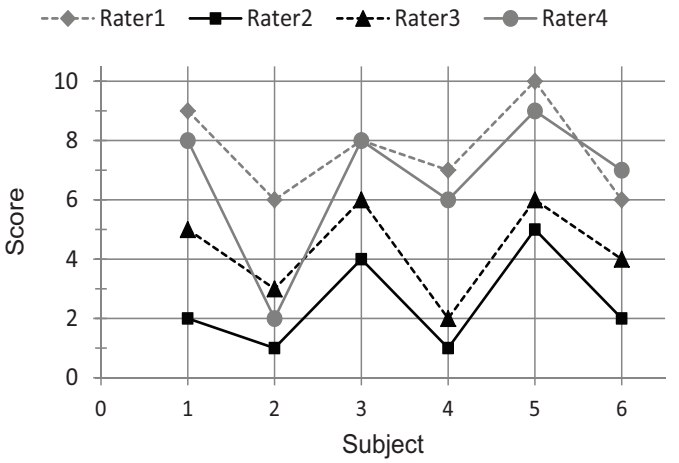

Figure 5.1: Ratings of 6 subjects by rater

Two types of factorial designs involving the subject and rater factors are the random and mixed factorial designs. The random factorial design is a design where the rater and the subject effects are random, while the mixed factorial design is one where the rater effect is fixed and the subject effect random.

In the random factorial design, the raters participating in the experiment are selected randomly from a larger universe of raters, and the participating subjects are selected randomly from a larger universe of subjects. The subjects and raters in their respective universes are actually those the researcher wants to investigate in the first place. The samples representing subgroups of these universes are used to minimize the costs of conducting experiments. It is the desire to draw meaningful conclusions about entire universes from their smaller representative samples that creates the need

to use statistical methods.

In the mixed factorial design on the other hand, only participating subjects are selected randomly from a larger subject universe. The participating raters are not tied to any other group of raters. They represent themselves, and are the only ones being investigated by the researcher. The study findings will only apply to these raters, and cannot be generalized to raters who did not participate in the experiment. For example, consider a reliability experiment whose purpose is to evaluate the consistency level between two measuring devices used in rheumatology clinical examinations. The researcher in this case, will want the study findings to be limited to the two specific measuring devices being investigated, and not be generalized to other devices that may not be similar to those used in the experiment. Experiments based on mixed factorial designs will often yield a higher ICC than those based on the random factorial designs, because no variation is generated by the rater effect when the design is mixed.

In this chapter, I will focus on the statistical methods used for analyzing experimental data based on the random factorial design. Methods needed for analyzing mixed factorial designs will be discussed in the next chapter.

5.2 The Intraclass Correlation Coefficients

The random factorial design involves a single group of raters as well as a single group of subjects, all of which are rated by each rater. That is the rater and subject factors are fully crossed. Table 5.1 shows lung functions data of 15 children representing their peak expiratory flow rates. Four measurements were taken on each subject by 4 raters. The raters here could represent 4 individuals operating the same measuring device, or one individual using the same measuring device on 4 different occasions. The data produced in these two scenarios can be analyzed with the same methods discussed in this section, although the results may be interpreted differently depending on the context.

Table 5.1 data were generated by a single group of 4 raters, each of whom rated all members of the same group of 15 subjects. We assume that the 4 raters are representative of a larger pool of raters they were selected from. Likewise, the 15 children are assumed to represent the larger population of children of interest they were randomly selected from.

Table 5.1: 15 children lung function measurements representing the peak expiratory flow rates (PEFR)[a]

Subject	Rater (j)			
(i)	1	2	3	4
1	190	220	200	200
2	220	200	240	230
3	260	260	240	280
4	210	300	280	265
5	270	265	280	270
6	280	280	270	275
7	260	280	280	300
8	275	275	275	305
9	280	290	300	290
10	320	290	300	290
11	300	300	310	300
12	270	250	330	370
13	320	330	330	330
14	335	320	335	375
15	350	320	340	365

[a]Source: Bland MJ, Altman DG. Statistics Notes: Measurement error. British Medical Journal 1996;312:1654 (extract)

Table 5.1 scores are described mathematically as follows:

$$y_{ijk} = \mu + s_i + r_j + (sr)_{ij} + e_{ijk}, \tag{5.2.1}$$

where y_{ijk} is the score assigned to subject i by rater j on the k^{th} trial[3]. The remaining terms of model 5.2.1 are defined as follows:

- μ is the expected value of the y-score for all subjects and raters.
- s_i is the random subject effect, assumed to follow the Normal distribution with 0 mean, and a variance σ_s^2. Moreover, $\text{cov}(s_i, s_{i'}) = 0$, if $i \neq i'$. That is the random subject effects are pairwise independent.
- r_j is the random rater effect, assumed to follow the Normal distribution with mean 0, and variance σ_r^2. These rater effects are pairwise independent. That is, $\text{cov}(r_j, r_{j'}) = 0$, if $j \neq j'$.

[3]Many reliability experiments only involves one trial (the first one)

- $(sr)_{ij}$ is the random subject-rater interaction effect, assumed to follow the Normal distribution with mean 0, and variance σ^2_{sr}. The interaction effects are assumed to be pairwise independent, with $\text{cov}\big((sr)_{ij}, (sr)_{i'j'}\big) = 0$, if $(i,j) \neq (i',j')$.
- e_{ijk} is the random error effect, assumed to follow the Normal distribution with mean 0, and variance σ^2_e.
- The subject, rater, and interaction effects are considered to be mutually independent. That is, the magnitude of one does not affect that of another effect.

It is essential for researchers to have a good understanding of the practical implications of some of these assumptions. Let us assume that the reliability experiment being analyzed involves n subjects, r raters, and m replicates (or trials). The fact that all rater effects (i.e. the r_j factors for raters $j = 1, \cdots, r$) share the same mean and the same variance σ^2_r indicates that all these raters have a similar understanding of the rating processes with their differences being random. If one rater systematically assigns high ratings to subjects, while a second rater assigns very low ratings to the same subjects, then the analysis of these ratings with model 5.2.1 may not be conclusive. This model will make the error term absorb most of the unexplained variation in ratings, which will result in low inter-rater and intra-rater reliability coefficients. Consequently, the ratings being analyzed must come from raters with a common understanding of the rating processes, which can be acquired with basic training.

Model 5.2.1, also referred to as Model 2 in the inter-rater reliability literature (see Shrout and Fleiss, 1979), stipulates that under the random factorial design, the different effects are additive, independent, and follow the Normal distribution. Unlike model 1A and 1B of the previous chapter, Model 5.2.1 allows for the calculation of both the inter-rater, and intra-rater reliability coefficients. I will review each of these coefficients in the next few sub-sections.

5.2.1 *Inter-Rater Reliability Coefficient*

An inter-rater reliability based on model 5.2.1 is by definition the correlation coefficient between the scores y_{ijk} and $y_{ij'k}$ associated with two raters j and j', the same subject i, and the same trial number k. It follows from equation 5.2.1 that the inter-rater reliability (denoted by ρ) is defined[4] as,

$$\boxed{\rho = \frac{\sigma^2_s}{\sigma^2_s + \sigma^2_r + \sigma^2_{sr} + \sigma^2_e}} \tag{5.2.2}$$

[4]Note that $\rho = \text{Corr}(y_{ijk}, y_{ij'k}) = \text{Cov}(y_{ijk}, y_{ij'k}) / \big[\sqrt{\text{Var}(y_{ijk})}\sqrt{\text{Var}(y_{ij'k})}\big]$

The question to be asked at this stage is whether equation 5.2.2 actually measures the extent of agreement among the r raters that participated in the experiment. A carefully examination of expression 5.2.2 suggests that ρ varies from 0 to 1, and takes a high value closer to 1 only when the subject variance σ_s^2 exceeds the combined variance $\sigma_r^2+\sigma_{sr}^2+\sigma_e^2$ by a wide margin. This will happen when the sum $\sigma_r^2+\sigma_{sr}^2+\sigma_e^2$ is small, which in turn indicates that the rater variance σ_r^2 is small. And a small rater variance σ_r^2 is a clear indication of high agreement among raters.

If a large value of ρ is a strong indication of good inter-rater agreement, can we say that a good inter-rater agreement will also result in a high value for ρ? The answer is unfortunately *"not necessarily."* In reality, a good inter-rater agreement will result in a high value for ρ only if the experiment is sufficiently well designed so as to keep the experimental error to the minimum. Again, it follows from equation 5.2.2 that a large error variance σ_e^2 will bring the whole ICC expression down even if the rater variance is small. Consequently,

> *if the ICC yields a high value, you can be certain that the extent of agreement among raters is good. However, if the ICC value is low, it is not necessarily an indication of poor agreement. It could be an indication of a poorly-designed experiment, and you may need to conduct additional analyzes.*

The experimental error may become unduly large, if your experiment is conducted in such a way that there are many uncontrolled factors that affect the magnitude of the scores other than the subject and the rater. These uncontrolled factors could be the location of the subject, major changes in experimental conditions such as the temperature, the measuring equipment or others. If the primary study objective is to obtain the ICC then the experimenter will want to design the experiment so that the subject and the rater are the most influential factors on the score magnitude. Adding more factors (controlled or uncontrolled) will negatively affect the ICC.

Calculating Inter-Rater Reliability

To compute the intraclass correlation coefficient from actual data, I again propose, a method that can handle missing scores, and which is an adaptation of what Searle (1997, page 474) has recommended. This method is more general than that proposed by Shrout and Fleiss (1979) , which cannot handle missing values. But both approaches match perfectly well when the data do not contain missing values.

The ICC of equation 5.2.2 is estimated from raw experimental data by calculating the 4 variance components σ_s^2, σ_r^2, σ_{rs}^2, and σ_e^2. The calculated subject, rater, rater-subject interaction, and error variances are respectively denoted by $\widehat{\sigma}_s^2$, $\widehat{\sigma}_r^2$, $\widehat{\sigma}_{rs}^2$, and $\widehat{\sigma}_e^2$. The calculated intraclass correlation coefficient is denoted by ICC(2,1)[5] and given

[5]The notation ICC(2,1) is widely used in the inter-rater reliability literature. ICC stands for

by:

$$\boxed{\text{ICC}(2,1) = \frac{\widehat{\sigma}_s^2}{\widehat{\sigma}_s^2 + \widehat{\sigma}_r^2 + \widehat{\sigma}_{rs}^2 + \widehat{\sigma}_e^2}.} \tag{5.2.3}$$

When there is no missing ratings, these variance components are conveniently calculated using a number of means of squares as shown in Shrout and Fleiss (1979) or McGraw and Wong (1996). The Shrout-Fleiss and McGraw-Wong procedures are discussed later in this section. Here, I like to first describe one method for calculating the variance components that accounts for missing ratings. This method is known in the statistical literature as the "Henderson Method I." There are several approaches for handling missing ratings that are proposed in the literature. My choice of the Henderson Method I is essentially motivated by its simplicity. The variance components are calculated as follows:

$$\widehat{\sigma}_e^2 = (T_{2y} - T_{2sr})/(M - \lambda_0), \tag{5.2.4}$$

$$\widehat{\sigma}_{sr}^2 = \Big[(M - k_1')\delta_r + (k_3 - k_2')\delta_s - \Big(T_{2s} - T_y'^2 - (n-1)\widehat{\sigma}_e^2\Big)\Big] \Big/ (M - k_1' - k_2' + k_5'), \tag{5.2.5}$$

$$\widehat{\sigma}_r^2 = \delta_s - \widehat{\sigma}_{sr}^2, \tag{5.2.6}$$

$$\widehat{\sigma}_s^2 = \delta_r - \widehat{\sigma}_{sr}^2, \tag{5.2.7}$$

where,

- T_{2y} is the summation of all squared values y_{ijk}^2.
- T_{2sr} is the summation of all factors $y_{ij.}^2/m_{ij}$, with $y_{ij.}$ being the total score value over all replicates associated with subject i and rater j, and m_{ij} the number of measurements taken on subject i by rater j.
- M is the total number of measurements produced by the experiment, while λ_0 is the number subject-rater combinations for which one measurement or more were produced (i.e. the number of (i,j)-cells for which $m_{ij} \geq 1$).
- $k_1' = k_1/M$, where k_1 is the summation of all $m_{i.}^2$ with $m_{i.}$ being the number of measurements taken on subject i.

Intraclass Correlation Coefficient, while "2" refers to model 2, and "1" indicates that each rating represents a single raw measurement, and not an average of several measurements.

- $\delta_r = \left[T_{2sr} - T_{2r} - (\lambda_0 - r)\widehat{\sigma}_e^2\right]/(M - k_4)$, where T_{2r} is obtained by summing all the terms $y_{\cdot j}^2/m_{\cdot j}$, with $y_{\cdot j}$ being the total score value associated with rater j. Moreover k_4 is calculated by summing all the terms $m_{ij}^2/m_{\cdot j}$, with $m_{\cdot j}$ being the number of measurements taken by rater j on all subjects.

- $\delta_s = \left[T_{2sr} - T_{2s} - (\lambda_0 - n)\widehat{\sigma}_e^2\right]/(M - k_3)$, where T_{2s} is obtained by summing all the terms $y_{i\cdot}^2/m_{i\cdot}$, with $y_{i\cdot}$ being the total score value associated with subject i. Moreover, k_3 is calculated by summing all the terms $m_{ij}^2/m_{i\cdot}$.

- $k_2' = k_2/M$, where k_2 is calculated by summing the terms $m_{\cdot j}^2$ over all raters.

- $T_y'^2 = T_y^2/M$, where T_y is the summation of all scores y_{ijk}, and $k_5' = k_5/M$ with k_5 calculated by summing the terms m_{ij}^2.

The calculation of the inter-rater reliability is described here for the factorial model with interaction effect. Such a model requires repeated measurements to be taken on the same subject. This is not always feasible, particularly if taking one measurement is very demanding on subjects. In this case, the researcher may prefer the use of a simpler model based on a single rating per subject and per rater, and in which the error and interaction effects are blended together. The calculation of the inter-rater reliability coefficient under this simpler model is discussed in section 5.5.1.

The following example illustrates how the intraclass correlation coefficient ICC(2,1) is calculated under model 5.2.1.

Example 5.1

For the purpose of this example, I have made two modifications to Table 5.1 to obtain Table 5.2 that will be analyzed. First, the number of children is reduced to 8, and two or three measurements were taken on some of the children. Second, the number of ratings per subject may vary from one rater to another creating some missing ratings as shown in Table 5.2.

Table 5.2 also contains some subject-level statistics such as the subject mean score, the subject standard deviation, and the mean score difference[6]. Figure 5.2 shows the graph of the subject mean score difference (c.f. column 8 of Table 5.2) against the subject mean score (c.f. column 6 of Table 5.2), and provides a visual assessment of the extent of agreement among the four raters. This graph can be seen as a version of what is known in the literature as the Bland-Altman plot - see Bland and Altman (1986). It follows from this graph that the 4 raters agree reasonably well with a possible exception

[6]The mean score difference is calculated separately for each subject by averaging all six pairwise differences associated with the six pairs of raters that can be formed out of the group of four raters.

of an outlier associated with a mean score of 303.75 and a mean score difference of 32.5.

Figure 5.3 on the other hand, depicts the subject standard deviation[7] as a function of their overall mean score. It offers another look at the rater agreement as the score magnitude changes, and is know in the literature as the repeatability plot (see Bland and Altman, 1996). This plot is often used for repeated measures, and can identify more outliers that the first plot would not. Both figures 5.2 and 5.3 tell about the same story showing an overall good agreement with the exception of 1 or 2 outliers.

The intraclass correlation coefficient ICC(2,1) of equation 5.2.3 is given by,

$$\begin{aligned} \text{ICC}(2,1) &= \frac{1,627.395}{1,627.395 + 82.507 + 0 + 460.897}, \\ &= 0.7497, \end{aligned}$$

where $\widehat{\sigma}_s = 1,627.395$, $\widehat{\sigma}_r = 82.507$, and $\widehat{\sigma}_e = 460.897$. As for the interaction variance component $\widehat{\sigma}_{rt}$, it was initially estimated to be a negative value -97.55, which was later replaced by 0 since the variance cannot be negative. The negative value obtained here is likely due to insufficient data for an accurate estimation of the interaction effect.

To see the details of these calculations, interested readers may look at the "Example 5.1" worksheet in the Excel spreadsheet,

`www.agreestat.com/books/icc5/chapter5/chapter5examples.xlsx`,

which shows the step-by-step calculations of ICC(2,1) for this example, from the input data of Table 5.2 to the final result of 0.7497.

[7] This standard deviation is that of all scores associated with one subject.

Table 5.2: Eight children lung function measurements representing the peak expiratory flow rates (PEFR)

Subject	Rater (j)					Standard	Mean
(i)	1	2	3	4	Mean Score	Deviation	Difference
1	190	220	200	200			
1	220	200	240	230	212.50	17.53	6.67
2	260	260	240	280			
2	210	300	280	265	261.88	27.51	15.42
3	270	265	280	270			
3	280	280	270	275			
3	260		280	300	275.45	10.60	6.53
4	275	275	275		275.00	0.00	0.00
5	280	290	300	290			
5	320	290	300	290	295.00	11.95	-3.33
6	300	300	310	300			
6	270	250	330	370	303.75	36.23	32.50
7	320	330	330	330			
7		320	335	375	334.29	18.80	17.50
8	350	320	340	365	343.75	18.87	10.83

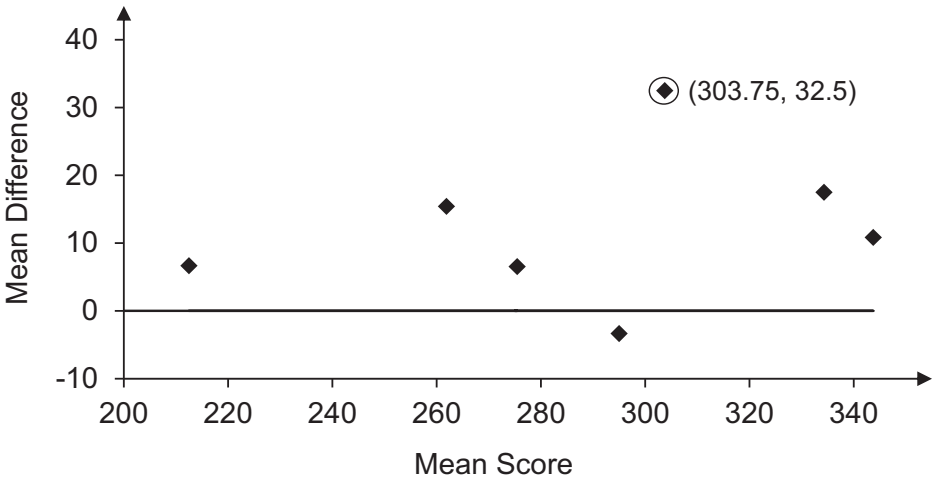

Figure 5.2: Mean score difference versus mean score based on Table 5.2 data.

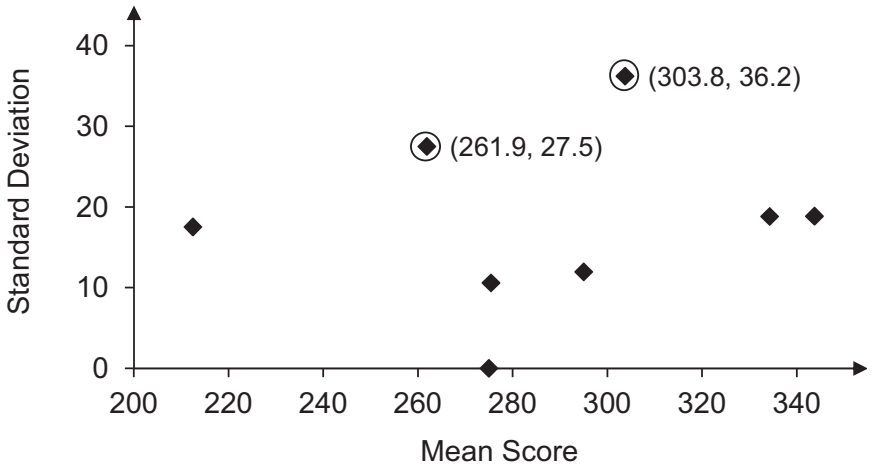

Figure 5.3: Repeatability plot showing the standard deviation against the mean score based on Table 5.2 data.

Some Simplifications with Complete Rating Data

When your rating data is complete, that is the measurements expected from all raters have been recorded, then the intraclass correlation ICC$(2,1)$ of equation 5.2.3 can take its usual and better known form given by,

$$\text{ICC}(2,1) = \frac{\text{MSS} - \text{MSI}}{\text{MSS} + r(\text{MSR} - \text{MSI})/n + (r-1)\text{MSI} + r(m-1)\text{MSE}}, \tag{5.2.8}$$

where MSS, MSR, MSI and MSE are defined as follows:

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

$$\text{MSS} = \frac{rm}{n-1}\sum_{i=1}^{n}(\overline{y}_{i\cdot\cdot} - \overline{y})^2 \tag{5.2.9}$$

- MSR is the mean of squares for raters, calculated by summing the squared differences $(\overline{y}_{\cdot j\cdot} - \overline{y})^2$, and by multiplying the summation by $nm/(r-1)$. The

term $\overline{y}_{.j.}$ represents the average of all ratings associated with rater j.

$$\text{MSR} = \frac{nm}{r-1}\sum_{j=1}^{r}(\overline{y}_{.j.} - \overline{y})^2. \tag{5.2.10}$$

- MSI is the mean of squares for the rater-subject interaction, calculated by summing the squared differences $(\overline{y}_{ij.} - \overline{y}_{i..} - \overline{y}_{.j.} + \overline{y})^2$, and by multiplying the summation by $m/[(r-1)(n-1)]$. The term $\overline{y}_{ij.}$ represents the average of all ratings associated with subject i and rater j.

$$\text{MSI} = \frac{m}{(r-1)(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{r}(\overline{y}_{ij.} - \overline{y}_{i..} - \overline{y}_{.j.} + \overline{y})^2. \tag{5.2.11}$$

- MSE is the mean of squares for errors, calculated by summing the squared differences $(y_{ijk} - \overline{y}_{ij.})^2$ and by dividing the summation by $rn(m-1)$.

$$\text{MSE} = \frac{1}{rn(m-1)}\sum_{i=1}^{n}\sum_{j=1}^{r}\sum_{k=1}^{m}(y_{ijk} - \overline{y})^2. \tag{5.2.12}$$

If your data is based on a single replication experimental design (i.e. only one measurement is taken on each subject), then the variances due to the error and the rater-subject interaction can no longer be calculated separately. In this case, only MSI must be calculated as described above, and be renamed as MSE. Equation 5.2.8 will then become,

$$\text{ICC}(2,1) = \frac{\text{MSS} - \text{MSE}}{\text{MSS} + r(\text{MSR} - \text{MSE})/n + (r-1)\text{MSE}}. \tag{5.2.13}$$

5.2.2 *Intra-Rater Reliability Coefficient*

The factorial design can be used for studying both the inter-rater and the intra-rater reliability. As previously discussed, intra-rater reliability is a measure of self-consistency for the raters. It represents the raters' ability to reproduce the same measurements on similar subjects. Therefore the intra-rater reliability study aims at investigating the reproducibility of the measurements. This task can be performed only if the same raters produce two ratings or more for the same subjects. That is, the notion of replication involving repeated trials plays a pivotal role in the study of intra-rater reliability.

The intra-rater reliability based on model 5.2.1, is by definition the correlation coefficient between the scores y_{ijk} and $y_{ijk'}$ associated with the two trials k and k', associated with the same subject i, and the same rater j. It follows from equation 5.2.1 that the intra-rater reliability (denoted by γ) is defined[8] as,

$$\boxed{\gamma = \frac{\sigma_r^2 + \sigma_s^2 + \sigma_{sr}^2}{\sigma_r^2 + \sigma_s^2 + \sigma_{sr}^2 + \sigma_e^2}.} \tag{5.2.14}$$

Does equation 5.2.14 actually measure raters' self-consistency? To answer this question, you should first note that γ would normally vary from 0 to 1, where 0 indicates no intra-rater reliability, and 1 a perfect intra-rater reliability. A high value for γ is an indication that the variance due to the error factor (σ_e^2) is relatively small compared to the total variance due to the rater, subject, and subject-rater interaction effects. Since the error variance is due to the combined effect of replication and experimental error, we can conclude that the variability due to replication is necessarily small. Therefore the ratings are highly reproducible. If γ is small (i.e. close to 0) then we can conclude that the combined effect of replication and experimental error is large. This may be due to a large variation in the repeated measurements (i.e. poor reproducibility) or a large experimental error or both. We don't know. What we know is that the experiment aimed at demonstrating high reproducibility was inconclusive.

Calculating Intra-Rater Reliability

To compute the intraclass correlation coefficient of equation 5.2.14 from actual data, I propose once again a method that can handle missing values adapted from Searle (1997, page 474). The method proposed by Shrout and Fleiss (1979), and which is based on several means of squares, cannot handle missing values. This special approach for balanced data, and the general approach for unbalanced data yield the same ICC when the data is balanced. For illustration purposes only, I will first present the special approach for balanced data.

SPECIAL METHOD FOR BALANCED DATA

If your data is balanced then the intra-rater reliability coefficient can be calculated as follows:

$$\text{ICC}_a(2,1) = \frac{r\text{MSR} + n\text{MSS} + (rn - r - n)\text{MSI} - rn\text{MSE}}{r\text{MSR} + n\text{MSS} + (rn - r - n)\text{MSI} + rn(m-1)\text{MSE}}, \tag{5.2.15}$$

[8] Note that $\gamma = \mathsf{Corr}(y_{ijk}, y_{ijk'}) = \mathsf{Cov}(y_{ijk}, y_{ijk'}) / \left[\sqrt{\mathsf{Var}(y_{ijk})}\sqrt{\mathsf{Var}(y_{ijk'})}\right]$. The assumption of independence of the factors can be used here to obtain equation 5.2.14

where r is the number of raters, n the number of subjects, m the number of trials (or replicates), MSR the mean of squares for raters, MSS the mean of squares for subjects, MSI the mean of squares for the subject-rater interaction, and MSE the mean of squares for errors. These different means of squares are calculated as shown towards the end of section 5.2.1 (after equation 5.2.8).

There is no need to know what this expression would be if there is only one measurement per rater and per subject. It is because the purpose of this section is to quantify reproducibility, which can be accomplished only in the context of replication involving two trials or more per subject.

GENERAL METHOD FOR UNBALANCED DATA

The ICC of equation 5.2.15 is calculated from raw experimental data by replacing the 4 variance components σ_s^2, σ_r^2, σ_{sr}^2, and σ_e^2 with their calculated values. The calculated subject, rater, interaction and error variances are respectively denoted by $\widehat{\sigma}_s^2$, $\widehat{\sigma}_r^2$, $\widehat{\sigma}_{sr}^2$ and $\widehat{\sigma}_e^2$. Therefore the calculated intraclass correlation coefficient for intra-rater reliability assessment is given by,

$$\boxed{\text{ICC}_a(2,1) = \frac{\widehat{\sigma}_r^2 + \widehat{\sigma}_s^2 + \widehat{\sigma}_{sr}^2}{\widehat{\sigma}_r^2 + \widehat{\sigma}_s^2 + \widehat{\sigma}_{sr}^2 + \widehat{\sigma}_e^2},} \tag{5.2.16}$$

where the variance components are calculated as shown in equations 5.2.4, 5.2.5, 5.2.6, and 5.2.7. Using Table 5.2 data and the variance components calculated in example 5.1, you can calculate the intra-rater reliability as,

$$\begin{aligned} \text{ICC}_a(2,1) &= \frac{1,627.395 + 82.507 + 0}{1,627.395 + 82.507 + 460.897}, \\ &= 0.788. \end{aligned} \tag{5.2.17}$$

It appears that the intra-rater reliability is reasonably high in this case. This is due to the fact that the error variance is small compared to the other variance components.

5.3 Statistical Inference About the ICC

The primary objective of this section is the present methods that allow you to quantify the precision of the intraclass correlation coefficient calculated with equations 5.2.3, and 5.2.16. You will be able to make a statement regarding the magnitude of the true[9] intraclass correlation coefficient using your experimental data. The process used to accomplish this task is known as statistical inference. The two inferential

[9]Recall that equations 5.2.3, or 5.2.16 can only give you an approximated value for the intraclass correlation based on the specific experimental data you have collected. The "true" ICC would require far more information about the entire population of subjects, than you can possibly collect.