CHAPTER 8

# Intraclass Correlations under the Random Factorial Design

OBJECTIVE

The objective in this chapter is to present methods for calculating various intraclass correlation coefficients and associated precision measures, in reliability studies where the rater and subject factors are fully crossed. Each rater is expected to rate all participating subjects, and may take a variable number of measurements on each of them. The rater and subject samples are both assumed to have been randomly selected from larger rater and subject populations, which the researcher is primarily interested in. Some intraclass correlations will be used for evaluating inter-rater reliability, while others will evaluate intra-rater reliability as a primary objective. Methods for obtaining confidence intervals and $p$-values will be presented as well.

CONTENTS

## 8.1 The Issues

The 2 one-factor models 1A and 1B presented in the previous chapter are simple to implement, but have inherent limitations that researchers must know. Model 1A produces an intraclass correlation coefficient that often underestimates the magnitude of the extent of agreement among raters. Likewise, the intraclass correlation coefficient calculated under model 1B tends to underestimate the magnitude of the intra-rater reliability (i.e. the extent to which each rater can reproduce measurements across similar subjects).

The ICC that is 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 3 effects, which are the rater effect, a possible rater-subject interaction effect[1], and the experimental error effect. Because these 3 effects are blended together, they are dependent to one another, and their combined variance will be higher than if the experiment was designed so that all 3 effects remain 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 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 this context 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 effects will often be high, reducing thereby the magnitude the ICC, and that of the intra-rater reliability. If an experiment is designed so that the rater, rater-subject interaction, and error effects are independent from one another, then it will achieve a smaller variance for these 3 effects, and 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.

[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)

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 there are 2 measurements or more that are taken on 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 inter-rater and intra-rater reliability. 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 8.1 depicts the reliability data of Table 7.1 of chapter 7 by subject. 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.

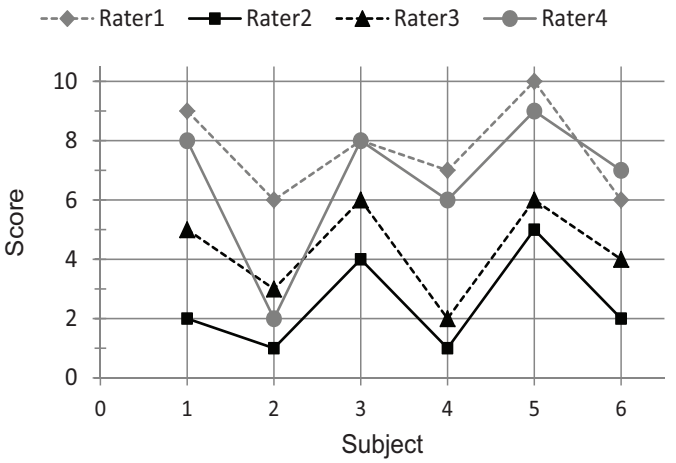


**Figure 8.1 :** Ratings of 6 subjects by rater

There are two types of factorial designs involving the subject and rater factors, which 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 on the other hand 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 2 measuring devices used in rheumatology clinical examinations. The researcher in this case, will want the study findings to be limited to the 2 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, we 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.

## 8.2 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 8.1 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 all these situations can be analyzed with the same methods that will be discussed in this section, although the results may be interpreted differently depending on the context.

Table 8.1 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 8.1**: 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)

The resulting scores are described mathematically as,

$$y_{ijk} = \mu + s_i + r_j + (sr)_{ij} + e_{ijk}, \tag{8.1}$$

where $y_{ijk}$ is the score assigned to subject $i$ by rater $j$ from the $k^{th}$ replication[3]. The remaining terms of model 8.1 are defined as follows:

- $\mu$ is the overall 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 $\sigma_s^2$.
- $r_j$ is the random rater effect, assumed to follow the Normal distribution with mean 0, and variance $\sigma_r^2$.
- $(sr)_{ij}$ is the random subject-rater interaction effect, assumed to follow the Normal distribution with mean 0, and variance $\sigma_{sr}^2$.

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

- $e_{ijk}$ is the random error effect, assumed to follow the Normal distribution with mean 0, and variance $\sigma_e^2$.

Moreover, 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. We will also assume that the reliability experiment involves $n$ subjects, $r$ raters, and $m$ replicates.

Model 8.1 (also referred to as Model 2 in the inter-rater reliability literature - see Shrout & 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 8.1 allows for the calculation of both the inter-rater, and intra-rater reliability coefficients. We will review each of these coefficients in the next few sub-sections.

### 8.2.1 *Inter-Rater Reliability Coefficient for Individual Ratings*

An inter-rater reliability based on model 8.1 is by definition the correlation coefficient between the scores $y_{ijk}$ and $y_{ij'k}$ associated with 2 raters $j$ and $j'$, the same subject $i$, and the same replicate number $k$. It follows from equation 8.1 that the inter-rater reliability (denoted by $\rho$) is defined[4] as,

$$\boxed{\rho = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_r^2 + \sigma_{sr}^2 + \sigma_e^2}} \tag{8.2}$$

The question to be asked at this stage is whether equation 8.2 actually measures the extent of agreement among the $r$ raters that participated in the experiment. A carefully examination of expression 8.2 suggests that $\rho$ varies from 0 to 1, and takes a high value closer to 1 only when the subject variance $\sigma_s^2$ 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 $\sigma_r^2$ is small. And a small rater variance $\sigma_r^2$ is a clear indication of high agreement among raters.

If a large value of $\rho$ 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 $\rho$? The answer is unfortunately *"not necessarily."* In reality, a good inter-rater agreement will result in a high value for $\rho$ only if the experiment is sufficiently well designed so as to keep the experimental error to the minimum. Again, it follows from equation 8.2 that a large error variance $\sigma_e^2$ will bring the who ICC expression down even if the rater variance is small. Consequently,

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