CHAPTER 4

Analysis of Quantitative Ratings

Contents

4.1 *Introduction* 94
4.2 *Intraclass Correlation: An Overview* 95
4.3 *Using the INTRACC.SAS Macro* 98
4.4 *The SAS/IML Function Modules for ICC* 102
4.4.1 *One-Factor ANOVA Models* 103
4.4.2 *Two-Factor Random ANOVA Models* 109
4.4.3 *Two-Factor Mixed ANOVA Models* 115
4.5 *Finn Coefficient* 121
4.5.1 *The Problem* 121
4.5.2 *The Solution* 123

4.1 Introduction

Chapters 2 and 3 are devoted to inter-rater reliability assessment for categorical ratings. The methods discussed in these 2 chapters are different versions of chance-corrected agreement coefficients and do not apply to quantitative ratings. A fixed and predetermined set of categorical ratings is generally made available to all raters before the beginning of the inter-rater reliability experiment, whereas quantitative ratings are produced by individual raters during the rating process. Since the magnitude of quantitative ratings is known only after the experiment had taken place, the extent of agreement among raters is generally quantified with the Intraclass Correlation Coefficient (ICC). This chapter discusses how ICC can be computed with SAS .

An influential paper in the field of ICC as a measure of inter-rater reliability was produced by Shrout and Fleiss (1979) . These authors proposed 6 ICC statistics that are often used by researchers in various fields of research. SAS built-in procedures will not produce these 6 statistics as specified by the authors. However, a SAS macro named `INTRACC.SAS` and often used by researchers can be downloaded using the following link:

https://support.sas.com/kb/25/addl/fusion25031_1_intracc.sas.txt.

The use of this macro is reviewed in section 4.3. It is a well documented program and you can find more details about its use with the following link:

https://support.sas.com/kb/25/031.html.

Note that the `INTRACC.SAS` macro is solely based on the work of Shrout and Fleiss (1979), which has a few important limitations:

- The methods discussed by Shrout and Fleiss (1979) cannot deal with missing ratings, which are common in the field of inter-rater reliability.

- Shrout-Fleiss methods are limited to experiments where a single rating is associated with each subject. However, experiments where subjects are rated multiple times by the same rater are common in practice. The methods for computing ICC estimates must be sufficiently general to include these scenarios.

To deal with these problems, I developed 2 SAS/IML libraries of function modules named `icc1factor.sas` and `icc2x3.sas`, which can compute the ICC$(1,1)$, ICC$(2,1)$ and ICC$(3,1)$ statistics regardless of the number of measurements per subject and whether there are missing ratings or not. These libraries can be downloaded using the following links:

https://agreestat.com/books/sas2/chap4/icc1factor.sas,
https://agreestat.com/books/sas2/chap4/icc2x3.sas.

Section 4.4 provides a more detailed coverage of these 2 libraries and shows examples of their use.

4.2 Intraclass Correlation: An Overview

The analysis of quantitative measurements generally starts with a hypothetical statistical model that is assumed to describe your data reasonably well. Analysis of Variance (ANOVA) models are typically used to describe quantitative rating data. Shrout and Fleiss (1979) have defined the following 3 main ANOVA models[1]:

- *Model 1.* This is a one-factor ANOVA where the rating y_{ij} associated with subject i and rater j is assumed to be the sum of the expected score μ, the random i^{th} subject effect s_i and the random error effect e_{ij}. More formally, you have,

$$y_{ij} = \mu + s_i + e_{ij}, \tag{4.2.1}$$

[1]These models are discussed in great details by Gwet (2021b)

where s_i follows the Normal distribution with mean 0 and variance σ_s^2, and e_{ij} follows the Normal distribution with mean 0 and variance σ_e^2. For this model, Shrout and Fleiss (1979) formulated an intraclass correlation coefficient referred to as ICC$(1,1)$.

- *Model 2.* This is two-way random effect ANOVA model, where the rating y_{ij} associated with subject i and rater j is assumed to be the sum of the expected score μ, the random i^{th} subject effect s_i, the random rater effect r_j, the subject-rater interaction $(sr)_{ij}$, and the random error effect e_{ij}. More formally, we have,

$$y_{ij} = \mu + s_i + r_j + (sr)_{ij} + e_{ij}, \tag{4.2.2}$$

where s_i follows the Normal distribution with mean 0 and variance σ_s^2, the rater effect r_j follows the Normal distribution with mean 0 and variance σ_r^2, the interaction effect $(sr)_{ij}$ follows the Normal distribution with mean 0 and variance σ_{sr}^2. Finally the error effect follows the Normal distribution with mean 0 and variance σ_e^2. The subject and rater effects are pairwise independent. For this model, Shrout and Fleiss (1979) formulated an intraclass correlation coefficient referred to as ICC$(2,1)$.

- *Model 3.* This is two-way mixed effect ANOVA model, where the rating y_{ij} associated with subject i and rater j is assumed to be the sum of the expected score μ, the random i^{th} subject effect s_i, the fixed rater effect r_j, the subject-rater interaction $(sr)_{ij}$, and the random error effect e_{ij}. More formally, we have,

$$y_{ij} = \mu + s_i + r_j + (sr)_{ij} + e_{ij}, \tag{4.2.3}$$

where s_i follows the Normal distribution with mean 0 and variance σ_s^2, the rater effects r_j $(j = 1, \cdots, r)$ sum to 0, the interaction effect $(sr)_{ij}$ follows the Normal distribution with mean 0 and variance σ_{sr}^2 and sum to 0 for any given subject i. Finally the error effect follows the Normal

distribution with mean 0 and variance σ_e^2. For this model, Shrout and Fleiss (1979) formulated an intraclass correlation coefficient referred to as ICC$(3,1)$.

All 3 intraclass correlation coefficients ICC$(1,1)$, ICC$(2,1)$ and ICC$(3,1)$ are implemented in the SAS macro `INTRACC.SAS`. Shrout and Fleiss (1979) proposed 3 additional intraclass correlations named ICC$(1,k)$, ICC$(2,k)$ and ICC$(3,k)$ and which aim at quantifying the intraclass correlation based on the analysis of the mean of k ratings, as opposed to the analysis individual ratings. These k-ICC estimates are implemented in the `INTRACC.SAS` macro as well.

> *I do not recommend using the k-ICC coefficients and will not discuss them in details in this book. Their use does not appear to be based on any sound theoretical or substantive ground.*

Some researchers even have a tendency of using k-ICC coefficients for the sole purpose of obtaining larger ICC estimates and claiming to have obtained good reliability. Shrout and Fleiss (1979) indicate that "*More typically, an investigator decides to use a mean as a unit of analysis because the individual rating is too unreliable.*" If the individual rating is too unreliable, then it is what it is and this should be reported. The average rating cannot be used as a surrogate for the individual rating and still lead to a valid measure of the extent of agreement among raters. Even if the study objective is to evaluate teams of physicians - an example used by Shrout and Fleiss (1979) - rather than individual physicians, the mean rating will still not be the answer. One should build a team consensus instead, either by reconciling differences through discussions or by using the first principal component[2]. An average of independent individual ratings within a team cannot be seen as a team attribute and will not lead to a valid inter-team reliability.

[2]For more details on the use of principal component analysis in inter-rater reliability assessment, see (Gwet, 2021b, sectin 9.4)

4.3 Using the INTRACC.SAS Macro

Let us consider the small dataset of Table 4.1. The SAS program 4.1 reads this data and computes all intraclass correlation coefficients described by Shrout and Fleiss (1979). The outcome of this program is shown in Figures 4.2 and 4.1.

Although the input dataset is supplied to the program in the same way it is presented in Table 4.1 (c.f. lines `#10` through `#15`), the program must read it in such a way that the created SAS dataset will look like the table in Figure 4.2. This is the only format that the `INTRACC.SAS` macro can accept. The use of this format is likely a constraint imposed by the GLM procedure used in the macro. The main advantage of this format is to allow you to analyze several rating variables simultaneously. This feature can prove useful if the subjects are rated on 2 or 3 different characteristics.

Line `#22` is where the program includes the `INTRACC.SAS` macro. Make sure you specify the directory into which you downloaded the macro. For this program, this directory is `c:\kgwet\sas2\chap4`. In line `#23` the `INTRACC.SAS` macro is called with the minimum number of parameters. The rating dataset `sfdata`, the dependent variable `score` containing the ratings, the variable representing the target (or subject) named `subject`, and the rater variable called `judge`. The results produced by this program are shown in Figure 4.1. The 3 most important numbers on this output are the following:

- "Shrout-Fleiss reliability: single score," or $\text{ICC}(1,1) = 0.16574$.
- "Shrout-Fleiss reliability: random set," or $\text{ICC}(2,1) = 0.28976$.
- "Shrout-Fleiss reliability: fixed set," or $\text{ICC}(3,1) = 0.71484$.

Before deciding which of the 3 statistics to use, you need to refer to the assumptions that underly the associated ANOVA model. $\text{ICC}(3,1)$ for example will not be appropriate if you want your analysis to apply to raters beyond those