CHAPTER 7

Finn's Coefficient of Reliability

OBJECTIVE

The objective of this chapter is to present a special inter-rater reliability proposed by Finn (1970), as an alternative to the more traditional intraclass correlation coefficient (ICC). The ICC approach to inter-rater can indeed fail. This will often be the case when the subject effect on ratings is small. A small subject effect violates a fundamental and often overlooked assumption of the ICC approach to inter-rater reliability, which requires the subject population to be quite diverse. For all practical purposes, you will know that you might be in trouble if the variance of the subject means is reasonably close to the mean of the subject variances. This chapter will also discuss the calculation of p-values associated with Finn's coefficient.

Contents

7.1 The Problem

As attractive as it may appear to mathematicians, the statistical theory underlying the ICC approach to inter-rater reliability is based on assumptions that cannot always be satisfied. For example, a practitioner may have a single subject (not a diverse population of subjects) on which the raters' agreement must be evaluated. In this case, the subject variance will be 0. Therefore, the ICC-based strategy in this case will collapse. It requires the rater variance component to be substantially smaller than the subject variance component to produce a high inter-rater reliability. In other inter-rater reliability applications, different subjects may be rated on different scales. For example patients suffering from different types of pain (back pain, knee pain, neck pain, ...) may have to be classified into different categories depending of their specific pain. If the same group of raters must rate this diverse population of subjects on different measurement scales, then the traditional inter-rater reliability methods will not work.

Finn's coefficient originally introduced by Finn (1970), is also known in the literature as the *Within-Group Inter-Rater Reliability* and was discussed under this name by several authors such as James et al. (1984), O'Neill (2017) and Smith-Crowe et al. (2014). and . Although Finn's coefficient was initially used for integer scores, I will generalize it in this chapter to handle quantitative scores. Unlike other more traditional inter-rater reliability coefficients, Finn's coefficient does not require a large subject sample nor independent subjects that could be seen as a random sample from a larger subject population. This method is not built on any formal statistical model either. It is less a pure statistical procedure in a traditional sense, and more a statistically-assisted quantitative procedure, which I have found very useful. "Orthodox" statisticians may have an issue with Finn's procedure. But our goal is to resolve practical problems, which are relevant to practitioners.

Consider the rating data shown in Table 7.1. This data is from an inter-rater reliability where 4 judges rated each of 5 subjects on 3 different occasions. If the 4 raters who produced these ratings differ from subject to subject, then inter-rater reliability may be quantified using Model 1A discussed in chapter 4. In this case, the error and subject variance components are respectively given by $\widehat{\sigma}_e^2 = 0.27659$ and $\widehat{\sigma}_s^2 = 0.00520$, which leads to ICC = 0.01846. If only 4 raters produced all the ratings and are assumed to have been randomly selected from a larger population of raters, then inter-rater reliability may be quantified using model 2 (with or without interaction) discussed in chapter 5. In this case, and assuming no interaction the error, subject and rater variance components are respectively given by $\widehat{\sigma}_e^2 = 0.25679$, $\widehat{\sigma}_s^2 = 0.00685$, and $\widehat{\sigma}_r^2 = 0.02420$, which leads to ICC = 0.02380. Likewise, the assumption of mixed model of chapter 6 yields ICC = 0.02598.

Now, the very low magnitude of all these ICC estimates is highly suspicious, and

not consistent at all with the perception one gets after a visual exploration of the dataset. No matter which angle you look at these ratings from, you arrive at the conclusion that the 4 raters largely agree in the way they see the 5 subjects being examined. What is the problem then? As it turns out, the culprit is the subject variance component. Let us consider the estimates based on model 1A for example (estimates based on the other 2 models 2, and 3 are very similar). With an estimated value of 0.00520, the subject variance is simply too small compared to the variance component associated with the combined effect of raters and experimental errors, which is estimated at 0.27659. The ICC as a measure of reliability is based on the following logic: *If the total variation in ratings is largely due to the subjects, then the variation due to the raters is small and therefore the raters are in agreement.* However, this logic is not telling us what will happen if the subjects are not generating much variation to begin with. That is where it might fail us.

Note that all 3 models discussed in chapters 4, 5,and 6 assume the subject effect to be random. This means that the group of subjects used in the inter-rater reliability experiment is assumed to be a representative sample of a larger population of subjects. The ratings assigned to these subjects are expected to vary from one subject to another, and from one rater to another. Moreover, that variation is explained by the type of subjects being rated (it is the subject effect), by the extent to which the raters agree (it is the rater effect), and by the measurement errors or experimental errors or reading errors (it is the error effect). The contribution of measurement errors can be minimized when the experiment is well designed. In this case the two main contributing factors to the rating variation will remain the subjects and the raters. One can see at this stage that a very homogeneous subject population expected to minimize the subject contribution to total variation in ratings will make it nearly impossible to obtain a high ICC. A highly diverse subject population on the other hand will favor a higher ICC for the same extent of agreement among raters. *Consequently, before deciding to use an ICC-based approach to inter-rater reliability, the practitioner needs first and foremost to secure the availability of a highly diverse subject population. Otherwise, the only approach that may help of the Finn's approach discussed in this chapter.*

Finn's coefficient can also be used as an effective management tool for monitoring the performance of raters coding, or clinical abstraction projects. Since it offers more flexibility with respect to the number of subjects and type of measurement scales it can accommodate, Finn's coefficient can be an integral part of routine Quality Assurance (QA) program. The raters can be tested on a regular basis, initially on a small number of subjects before deciding whether testing on a larger scale is necessary to evaluate the magnitude of an identified problem.

Instead of comparing the rater variance to the subject variance to determine the inter-rater reliability coefficient as the ICC, Finn's coefficient compares the rater vari-

ance component to its expected value under the assumption of no reliability among raters. The observed rater variance could be smaller or higher than its expected under the no-reliability assumption, it is only when it is substantially higher would the raters be considered highly reliable.

Table 7.1: Ratings of 5 Subjects by 4 Judges

Subject	Judges J1	J2	J3	J4	Row Variance
1	6	5.6	5	5	
1	5	5	5	5	0.2118
1	5	5	5	4	
5	6	6	5	5	
5	5	5	5	5	0.1875
5	5	5	5	4.5	
4	6	6	5	5	
4	5	5	5	5	0.1515
4	5	5	5	5	
2	6	6	5	5	
2	5	5	5	5	0.1515
2	5	5	5	5	
3	6	6	5	5	
3	5	5	5	5	0.6806
3	3.5	3.8	3.6	4.3	
Average					0.2766

7.2 Finn's Coefficient

In its original version, Finn's coefficient discussed by Finn (1970), was described under the assumption that the ratings are sequential integer values $1, 2, 3, \cdots, q$ where q is the number of categories considered in the study. This original version is discussed in section 7.2.1. However, I extended this version to the more general set of q ratings $(x_1, \cdots, x_k, \cdots, x_q)$ where x_k is a real number that could be of a ratio or an interval type. This generalized version of Finn's coefficient will be discussed in section 7.2.2.

7.2.1 *Finn's Coefficient: Computations*

Table 7.3 used by Finn (1970) to illustrate his method, shows ratings that 5 judges assigned to 4 items. Each judge was expected to assign one of the integer values (1, 2, 3, 4, 5) to each of the 4 items. Finn's method can be summarized as follows:

- Compute all 4 row (or subject-level) variances[1]. For Table 7.3, the 4 subject variances are 0.2, 0.0, 0.20, and 0.20.
- Average all 4 subject-level variances to obtain the mean subject variance (MSV) of 0.15. This number tells us how far on average the rating from any given judge will stray away from the average rating. The smaller the mean subject variance, the higher the rater agreement.
- Compute the expected value that this mean subject variance would take if the ratings were assigned to subjects in a purely random manner. If q is the total number rating values the judges can use (in our case $q = 5$) then the mean subject variance under the assumption of random rating is $(q-1)(q+1)/12$ (for Table 7.3, this expected mean subject variance is $(5-1)(5+1)/12 = (4 \times 6)/12 = 2$.
- Compute Finn's coefficient as follows:

$$r_{\text{F}} = 1 - \frac{\textit{Observed}\ \text{MSV}}{\textit{Expected}\ \text{MSV}}. \tag{7.2.1}$$

To compute this coefficient, several variances are first calculated at the subject level. Even when several measurements are taken on each subject as in Table 7.1, variances must still first be computed within the group of measurements associated with one subject. Hence, the name "Within-group inter-rater reliability."

Using equation 7.2.1 and Table 7.3 data, Finn's coefficient is calculated as $r_{\text{F}} = 1 - 0.15/2 = 0.925$. You will see in subsequent chapters that the ICC applied to these ratings produces an unduly low inter-rater reliability coefficient.

Note that Finn (1970) assumed that the MSV reaches its maximum value if the ratings are assigned to subjects in a purely random manner. In reality, the observed MSV could well exceed its expected value under the assumption of random rating. Consider the rating data in Table 7.2, where 5 judges assigned

[1]Note that these are sample variances, which use 3 (i.e. $4-1$) in their denominators.

one of 5 scores (1, 2, 3, 4, 5) to the 5 subjects (A, B, C, D, E). The observed MSV equals 3.24, which exceeds its expected value of 2 by a wide margin. This would lead to a Finn's coefficient of $r_F = 1 - 3.24/2 = -0.620$. This itself is not a major problem, as any zero or negative Finn's coefficient can safely be interpreted as an absence of reliability.

Table 7.2: Ratings of 4 Items by 5 Judges with a Negative Finn's Coefficient

	Judges					
Subjects	I	II	III	IV	V	Row Variance
A	1	1	2	4	5	3.3
B	1	3	5	5	5	3.2
C	1	1	3	4	5	3.2
D	1	2	4	5	5	3.3
E	1	2	3	5	5	3.2
Average						3.24

Table 7.3: Ratings of 4 Items by 5 Judges with a Positive Finn's Coefficient

	Judges					
Items	I	II	III	IV	V	Row Variance
A	2	2	3	2	2	0.20
B	2	2	2	2	2	0.00
C	2	2	2	2	1	0.20
D	1	2	2	2	2	0.20
Average						0.15

7.2.2 *Finn's Coefficient: General Definition*

Finn (1970) introduced his coefficient with a special focus on categorical data. Although other methods by Cohen (1960), Fleiss (1971), Gwet (2008) are available in the literature for categorical data, Finn's coefficient may still be useful if the number of subjects is very small. Finn's coefficient can also be used with quantitative data as an alternative to the ICC when the number of subjects is small. In this section,

I will present a more general version of Finn's coefficient that could be used with categorical as well as for quantitative ratings.

Let us consider r raters who must assign one of q scores $(x_1, \cdots, x_q)$ to each of n items. The score associated with subject i and rater j will be denoted by $x^\star_{ij}$, which equals one of the q values $(x_1, \cdots, x_q)$. Only after the inter-rater reliability had been conducted, will you know which of the q scores equals $x^\star_{ij}$ for any particular subject i, and rater j. The mean subject variance is given by,

$$\mathrm{S}^2 = \frac{1}{n}\sum_{i=1}^{n} \mathrm{S}_i^2, \text{ where } \mathrm{S}_i^2 = \frac{1}{r-1}\sum_{j=1}^{r}\left(x^\star_{ij} - \overline{x}^\star_{i\cdot}\right)^2, \tag{7.2.2}$$

with $\overline{x}^\star_{i\cdot}$ being the average of the $x^\star_{ij}$ values associated with subject i, and S_i^2 subject i's variance. Now, if the q values $(x_1, \cdots, x_q)$ were to be assigned randomly to the n subjects then the expected mean subject variance S_E^2 (or Expected MSV) is given by,

$$\mathrm{S}_\mathrm{E}^2 = \frac{1}{q}\sum_{k=1}^{q}(x_k - \overline{x})^2. \tag{7.2.3}$$

Finn's coefficient is now defined as follows:

$$\boxed{r_\mathrm{F} = 1 - \mathrm{S}^2/\mathrm{S}_\mathrm{E}^2.} \tag{7.2.4}$$

Note that if the q ratings $(x_1, \cdots, x_q)$ used in the experiment are the first q integer values $(1, 2, \cdots, q)$, then equation 7.2.4 becomes,

$$\boxed{r_\mathrm{F} = 1 - 12\mathrm{S}^2/\big[(q-1)(q+1)\big].} \tag{7.2.5}$$

If you want to account for missing values by assigning them a rank of 0, then the $q+1$ values 0, 1, $\cdots$, q will be available for use to each rater, and Finn's equation will become:

$$\boxed{r_\mathrm{F} = 1 - 12\mathrm{S}^2/\big[q(q+2)\big].} \tag{7.2.6}$$

Finn's generalized method is illustrated in the following Example 7.1:

Example 7.1

Let us again consider the ratings of Table 7.1 previously discussed in section 7.1. To compute Finn's inter-rater reliability coefficient, one needs to first compute the 5 subject-level variances (or within-subject variances) shown in the rightmost column of Table 7.1. The next step is to average all 5 subject-level variances to obtain the mean subject variance (MSV) of 0.27659.