CHAPTER 8

Measures of Association and Concordance

OBJECTIVE

The main objective of this chapter is to review some special agreement coefficients that are often used to quantify the extent of agreement among raters, with quantitative ratings. These agreement coefficients, which are different from the intraclass correlation coefficients discussed in part II of this book, are sometimes preferred for their simplicity. They can be used with quantitative ratings without appealing to any Analysis Of Variance model. I will discuss some of their advantages as well as their limitations.

Contents

8.1 Overview

Throughout this chapter, I assume that you are dealing with quantitative ratings from 2 raters or more, and which are to be used for quantifying the extent of agreement. We have learned from the previous chapters that with quantitative ratings, you would normally use one version of the intraclass correlation coefficient (ICC). However, the use of one of these ICC versions requires that you first describe the Analysis Of Variance (ANOVA) model the underlies your data. The methods presented in this chapter provide a close formula that can be readily applied to your data to obtain the desired agreement coefficient. However, they also present some limitations that you must be aware of.

When dealing with 2 raters only, one option consists of calculating the traditional Pearson's correlation coefficient. You will see that while a high agreement inevitably leads to a high Pearson's correlation coefficient, the opposite is not necessarily true. Consequently, a high Pearson's correlation must be carefully interpreted before a definitive conclusion can be made. This represents the biggest issue with using Pearson's correlation as a measure of agreement. These issues and others are discussed in section 8.2.1.

The Pearson's correlation is a parametric method in the sense that any statistical inference (e.g. p-value, confidence intervals, ...) requires that an assumption be made about the probability distribution of the ratings. This requirement can be avoided with the Spearman's correlation coefficient to be discussed in section 8.2.2. This is a pure nonparametric method based on rating ranks and that requires no assumption about the probability distribution of the ratings. It works well for 2 raters, although it may carry the traditional problems associated with rank-based statistics, which their lack of sensitivity to small changes in the raw ratings.

Similar to Spearman's correlation, Kendall's Tau is another nonparametric correlation coefficient that has been used as a measure of agreement between 2 raters. The only I could find why some researchers would use Kendall's Tau as opposed to Spearman's correlation is that the former appears some interesting statistical properties such unbiasedness, which may not be of critical importance to most practitioners. Kendall's Tau will be discussed in section 8.3.

In the group of nonparametric agreement coefficient, I will also discuss Kendall's Coefficient of Concordance (KCC) in section 8.4. It is the only nonparametric method discussed in this book, which can quantify the extent of agreement among 3 raters or more. This method may be somehow problematic if it is to be used for comparing 2 studies, as it may be able to capture small differences effectively. Otherwise, I deem this approach to be useful for researchers who may not want to use the intraclass correlation coefficient.

The last method discussed in this chapter (section 8.5) is Lin's concordance correlation coefficient (LCCC). It can only be applied to 2 raters and is a parametric method. You may remember that a high Pearson's correlation may not necessarily be a reliable indicator of high agreement. Lin's coefficient aims at correcting this deficiency. It is indeed similar to Pearson's correlation even though it requires more complex computations. In my opinion, this is an interesting option for researchers to consider if the intraclass correlation coefficient cannot be used.

8.2 Pearson & Spearman Correlation Coefficients

In this section, I present two related bivariate measures of association named the Pearson Product-Moment Correlation Coefficient (better known as Pearson Correlation), and the Spearman's Rank-Order Correlation Coefficient (better known as Spearman Correlation). Each of these two measures can only evaluate the extent to which ratings from two raters are related. The Pearson correlation requires stringent conditions to be met to ensure its validity. When these conditions are not met, Spearman's correlation is often the alternative of choice.

Table 8.1 is an extract of Table 5.1 of chapter 5, and represents lung function measurements on 15 children produced by Rater 1 and Rater 2. The relationship between the two series of ratings is depicted in Figure 8.1, where one may see a linear trend. This data will be used to illustrate the calculation of the two correlation coefficients discussed in this section. Let r designate the Pearson correlation coefficient, and r_s the Spearman's correlation. Let X_1 and X_2 represent the abstract ratings associated with raters 1 and 2 respectively.

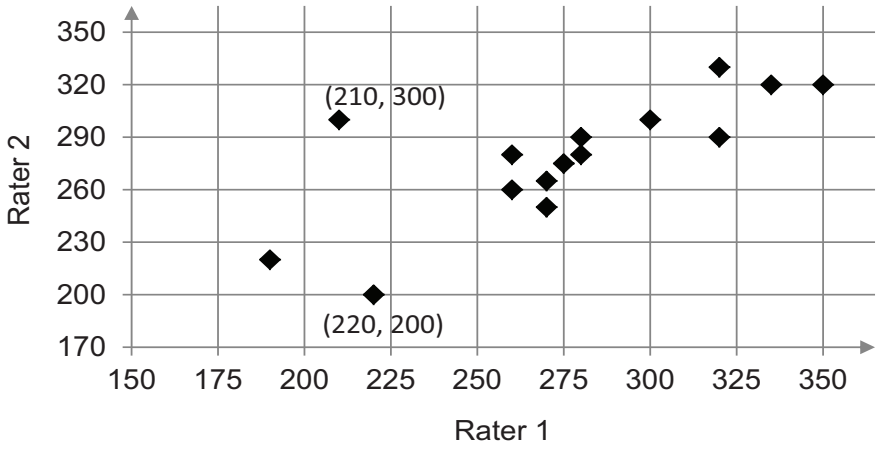

Figure 8.1: Lung function measurements on 15 children by Rater 1 and Rater 2.

Table 8.1: 15 children lung function measurements representing the peak expiratory flow rates (PEFR), and taken by raters 1 and 2

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

8.2.1 *Pearson's Correlation Coefficient*

The Pearson's correlation is one of the best known coefficients in statistical science, and has been widely used across many fields of research since it was introduced by Pearson (1896, 1900). It is calculated as follows:

$$\boxed{r = \frac{cov(X_1, X_2)}{\sqrt{v(X_1)}\sqrt{v(X_2)}},} \tag{8.2.1}$$

where $cov(X_1, X_2)$ is the covariance between the ratings X_1 and X_2 associated with raters 1 and 2 respectively, while $v(X_1)$ and $v(X_2)$ are the variances. The variance $v(X_1)$ is a statistical measure that tells you how far you can expect any given rating from rater 1 to stray from the overall average. The covariance $cov(X_1, X_2)$ on the other hand, is based on the differences of the ratings to their overall means for both raters 1 and 2. Covariance tells you how large you may expect the product of these differences to be. A positive covariance is an indication that both series of ratings change in the same direction. If negative, it indicates that the ratings change in opposite directions.

The correlation coefficient of equation 8.2.1 can be calculated with MS Excel using the function CORREL(). All standard statistical packages have functions for calculating the same coefficient. Using MS Excel, I calculated the correlation coefficient associated with Table 8.2.4 as follows:

$$r = \frac{1,242.5}{\sqrt{2,065.0}\sqrt{1,301.7}} = 0.758.$$

Note that r quantifies the linearity of the relationship between the 2 series of ratings. As such. it is often referred to as *linear correlation coefficient*. This coefficient may be used as a measure agreement under certain conditions. High agreement will always translate into a high Pearson's correlation. However, a high Pearson's correlation will not necessarily be associated with high agreement[1]. Therefore, a low

[1]Pearson's correlation may still be high even though one series of ratings is systematically twice higher than the other (a situation that I must admit, is very uncommon for rating data).

Pearson's correlation is an accurate indication of low agreement. It is when this correlation is high that you should not jump into the conclusion that you have high agreement prior to verifying that both series of ratings can be seen as coming from the same distribution with a common mean (a classical t-test can be used for this purpose). Consequently, this coefficient cannot be blindly used as a measure of agreement.

The main interest in Pearson's correlation stems from its simplicity and its popularity in the statistical community. Everybody know about it. However, practitioners are not always aware that the validity of Pearson's correlation coefficient requires a specific list of conditions to be satisfied. Here are these conditions:

(a) The sample of n subjects is randomly selected from the population it represents.

(b) The measurement level associated with each series of ratings is interval or ratio data.

(c) Each of the series of ratings follows the Normal distribution.

Other conditions irrelevant for rating data, are often required to ensure the validity of the Pearson's correlation coefficient.

Computing the p-Value for Pearson's Correlation Coefficient

Let t be a statistic defined as follows:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}, \tag{8.2.2}$$

where n is the number of subjects that participated in the reliability experiment, and r the Pearson's coefficient. The statistic t of equation 8.2.2 follows the Student's distribution[2] with $n-2$ degrees of freedom. The p-value, which measures the statistical significance of the Pearson's correlation coefficient can be calculated using MS Excel as follows:

"=T.DIST.2T(ABS(t),$n-1$)", *for Excel 2010/2011 or a more recent version,*
"=TDIST(ABS(t),$n-1$,2)", *for Excel 2007/2008 or an earlier version.*

Note that the notation Excel 2010/2011 refers to the 2010 Windows version and 2011 Mac version of Excel.

Because the last two validity conditions (b) and (c) associated with the Pearson's correlation coefficient are often violated, practitioners frequently use the Spearman's

[2]Readers who need to know more about the Student's law of probability may want to read an introductory statistics book.

correlation coefficient (also known in the literature as the Spearman's "Rank-Order" Correlation Coefficient).

8.2.2 *Spearman's Correlation Coefficient*

The Spearman's correlation coefficient is a bivariate measure of correlation, which requires rank-order input data, and which quantifies agreement between 2 sets of ratings. It is a nonparametric method solely based on the ranks associated with the original ratings, and not on the ratings themselves. That is, original ratings from both raters are initially lumped together and ranked as one set of ratings. The resulting ranks remain with the same subjects that are then reassigned to their original groups. The Spearman's correlation is calculated using the subject-level differences between rank values. This is a more reliable measure of agreement than Pearson's correlation coefficient with a downside. The downside comes from the use of ranks that makes this measure less sensitive to major changes in the magnitude of the ratings. That is, a noticeable change in the magnitude of raw ratings may not alter their respective rankings, leading to the exact same Spearman's correlation value.

Calculating Spearman's Correlation

The procedure for computing the Spearman's correlation is best described with an example. Table 8.2 shows the children lung function data previously analyzed in this section. The "Rank 1" row represents the rankings associated with Rater 1's ratings, while row "Rank 2" contains similar rankings for Rater 2. The rank differences (i.e. Rank 1 - Rank 2) and their squared values calculated in rows d and d^2 are used in the calculation of Spearman's correlation (denoted by r_{S}) as shown in the following equation:

$$r_{\text{S}} = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}, \tag{8.2.3}$$

where n is the number of subjects (in Table 8.2, $n = 15$), and d_i the rank difference associated with subject i. Based on the information shown in Table 8.2, the Spearman's correlation coefficient is obtained as,

$$r_{\text{S}} = 1 - \frac{6 \times 145}{15 \times (15^2 - 1)} = 0.741.$$

Validity Conditions

Spearman's correlation was developed by Spearman (1904), and can be used if one of the following conditions is satisfied:

(*a*) The data for both series of ratings are in a rank-order format.

(*b*) The original data are in a rank-order format for one set of ratings and in an interval/ratio for the second (this latter variable will then have to be converted to ranks before Spearman's correlation is calculated.

(*c*) The data for both series of ratings have been transformed into a rank-order format from an interval/ratio because the validity conditions associated with the Pearson's coefficient may not be satisfied.

The p-value associated with Spearman's correlation coefficient can be calculated the exact same way it is calculated for Pearson's correlation coefficient.

Table 8.2: Spearman's correlation calculation based on Table 8.1 children lung function measurements

Subject	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Total
Rater 1	190	220	*260*	210	**270**	***280***	*260*	275	***280***	**320**	300	**270**	**320**	335	350	
Rater 2	220	200	260	*300*	265	**280**	**280**	275	***290***	***290***	*300*	250	330	**320**	**320**	
Rank 1	1	3	4.5	2	6.5	9.5	4.5	8	9.5	12.5	11	6.5	12.5	14	15	
Rank 2	2	1	4	11.5	5	7.5	7.5	6	9.5	9.5	11.5	3	15	13.5	13.5	
d	-1	2	0.5	-9.5	1.5	2	-3	2	0	3	-0.5	3.5	-2.5	0.5	1.5	
d^2	1	4	0.25	90.25	2.25	4	9	4	0	9	0.25	12.25	6.25	0.25	2.25	145

Treatment of Ties

If one of the two series of ratings being analyzed contains ties (i.e. repeated ratings or repeated ranks) then it is recommended to compute a tie-adjusted Spearman correlation[3] (denoted by $r_{\text{S}}^{\star}$), which is given by,

$$r_{\text{S}}^{\star} = \frac{A_1 + A_2 - \sum_{i=1}^{n} d_i^2}{2\sqrt{A_1 A_2}}, \tag{8.2.4}$$

where A_1 and A_2 are now defined. A tie series is defined as a set of ranks that has identical values. It appears from Table 8.2 that each of the two raters has produced

[3]The need for tie adjustment comes from the fact that equation 8.2.3 tends to inflate artificially the absolute value of Spearman's correlation.

4 tie series containing two scores each. For example, the first tie series from Rater 1 contains the ranks $\{4.5, 4.5\}$ associated with the scores $\{260, 260\}$. Let i be a particular tie series and t_i the associated number of scores. Consider the following quantities:

- T_1 = summation of the differences $t_i^3 - t_i$ for all tie series produced by Rater 1, and $A_1 = (n^3 - n - T_1)/12$. Using Table 8.2.5 data, T_1 can be calculated as $T_1 = 4 \times (2^3 - 2) = 24$ (note that there are 4 tie series from Rater 1 that contains 2 scores each). Therefore, $A_1 = (15^3 - 15 - 24)/12 = 278$.

- T_2 = summation of the differences $t_i^3 - t_i$ for all tie series produced by Rater 2, and $A_2 = (n^3 - n - T_2)/12$. It follows from Table 8.2 that $T_2 = 4 \times (2^3 - 2) = 24$ (note that there are 4 tie series from Rater 2 that contains 2 scores each). Therefore, $A_2 = (15^3 - 15 - 24)/12 = 278$.

Consequently, the tie-adjusted Spearman correlation is calculated from equation 8.2.4 as follows:

$$r_S^\star = \frac{278 + 278 - 145}{2 \times \sqrt{278 \times 278}} = 0.7392.$$

Those interested in a more detailed discussion of tie-adjusted Spearman's correlation coefficient could find it in Sheskin (2004, pp. 1067-1069).

8.3 Kendall's Tau

Tau is a Greek character such as alpha, and its symbol τ is used to designate the bivariate measure of association proposed by Kendall (1938). This coefficient is discussed here because a number of researchers have used it to quantify the degree of agreement between rankings from two judges. Kendall's tau is a bivariate coefficient which, like Spearman's correlation, is based on rank-order data. The question then becomes, why Kendall tau when Spearman's correlation was proposed earlier? Although tau uses a more tedious computation procedure, it possesses interesting statistical properties such as unbiasedness that the Spearman's correlation lacks. Moreover, Lindeman et al. (1980) indicated that the Normal distribution provides a good approximation to the sampling distribution of tau for small sample sizes. These appear to be the main reasons some researchers prefer tau to Spearman's correlation.

Kendall's tau can be seen as a measure of the extent of agreement between two raters based on their ranking of subjects. The ranks associated with subjects and derived from the initial numeric ratings produced by the 2 raters, are compared not for each individual subject, but for each pair of subjects. That is, for a pair of subjects (i, j) one can determine their rankings (A_i, A_j), and (B_i, B_j) with respect to the two raters A and B. If the sign of the difference $A_i - A_j$ is the same as the sign of the difference $B_i - B_j$, then the pair (i, j) is said to be concordant, otherwise it is said

to be discordant. Kendall's tau is essentially the difference between the proportion of concordant pairs and the proportion of discordant pairs. In the next 2 sections, I will describe the procedure for computing Kendall's Tau coefficient, when the set of ratings contains no ties (i.e. duplicate ratings), and when it does.

8.3.1 *Computing Kendall's Tau in the Absence of Ties*

You will see that there is no tie in the two sets of ratings (and therefore in the associated rankings) when no pair of subjects has two identical values from either rater. In this case, any pair of subjects will be either concordant or discordant. The method for dealing with ties will be discussed in the next section.

Let n_C be the number of concordant pairs of subjects, and n_D the number of discordant pairs. If n is the total number of subjects that participated in the experiment, then Kendall's tau is calculated as follows:

$$\tau = \frac{n_C - n_D}{n(n-1)/2}. \tag{8.3.1}$$

Note that $n(n-1)/2$ is the total number of distinctive pairs of subjects. The following example illustrates the calculation of Kendall's tau using a small sample of 6 subjects, and equation 8.3.1.

Example 8.1

Table 8.3 shows 6 subjects and their raw ratings from two judges named Judge1 and Judge2. Rank1 and Rank2 represent the rankings in ascending order of Judge1's and Judge2's ratings. The calculation of Kendall's tau will be based entirely on these rankings as shown in Table 8.4.

Table 8.3: Ratings and Rankings of Six Subjects Scored by Judge 1 and Judge 2

Subject	Judge1	Judge2	Rank1	Rank2
1	9	2.7	5	4
2	6.6	1.4	2	2
3	8	4	4	5
4	7.1	1	3	1
5	10	5.8	6	6
6	6	2	1	3

The first row of Table 8.4 labeled as "Subject #" contains the subject names or labels, the "Rank1" and "Rank2" rows represent the rankings associated with the judges 1 and 2 respectively. The numbers on the table diagonal are Judge 2's rankings. The letters C and D indicate concordant and discordant pairs of subjects formed by the column and row labeled as "Subject #." The first letter C of row $\boxed{1}$ for example, is associated with the pair of subjects $(1, 2)$. This pair is concordant because the associated pairs of ranks (i.e. $(5, 2)$ from Judge1, and $(4, 2)$ from Judge2) change in the same direction (i.e. $5-2=3>0$ and $4-2=2>0$). The first D letter in that same row is associated with the pair of subjects $(1, 3)$. The 2 associated pairs of rankings vary in opposite direction. Therefore, $(1, 3)$ is a discordant pair of subjects. The remaining rows of the table are obtained in the same manner. For example the first C letter of row (2) is associated with the pair of subjects $(2, 3)$.

The symbols n_{C} and n_{D} represent the number of concordant and discordant pairs of subjects. It follows from the last row of Table 8.4 that the total number of concordant pairs is 11, while the number of discordant pairs is 4. Using equation 8.3.1, we can compute Kendall's tau as,

$$\tau = \frac{11-4}{6\times(6-1)/2} = \frac{7}{15} = 0.4667.$$

Table 8.4: Ratings and Rankings of Six Subjects Scored by Judge 1 and Judge 2

Subject #	$\boxed{1}$	$\boxed{2}$	$\boxed{3}$	$\boxed{4}$	$\boxed{5}$	$\boxed{6}$		
Rank1	5	2	4	3	6	1		
Rank2	4	2	5	1	6	3	n_{C}	n_{D}
Subject #								
$\boxed{1}$	4	C	D	C	C	C	4	1
$\boxed{2}$		2	C	D	C	D	2	2
$\boxed{3}$			5	C	C	C	3	0
$\boxed{4}$				1	C	D	1	1
$\boxed{5}$					6	C	1	0
$\boxed{6}$						3	0	0
Concordant & Discordant Pairs							11	4

8.3.2 *Computing Kendall's Tau in the Presence of Ties*

When the judges produce ties, then the Kendall's tau coefficient must be properly adjusted. The definition of concordant and discordant pairs of subjects is

incompatible with the existence of ties, since the rank difference associated with pair of subjects with ties will always be 0 (i.e. will have no direction). To see this, consider a pair of subjects (i, j) and the associated ranks $(3.5, 3.5)$ and $(1, 2)$ from 2 judges 1 and 2. This pair of subjects is neither concordant nor discordant since only the second set of ranks changes, the first one remaining unchanged. The tie-adjusted Kendall's tau allows you to exclude pairs with ties from the number of concordant and discordant pairs, and to adjust the denominator accordingly. This tie-adjusted tau coefficient $\tau^\star$ is calculated as follows:

$$\tau^\star = \frac{2(n_{\text{C}} - n_{\text{D}})}{\sqrt{n(n-1) - T_1} \times \sqrt{n(n-1) - T_2}}. \tag{8.3.2}$$

To define T_1 and T_2, let i be a specific set of ties, and t_i the number of subjects associated with that set. Then T_1 is the summation of all values $(t_i^3 - t_i)$ for all sets of ties produced by Judge 1. Likewise, T_2 is the summation of all values $(t_i^3 - t_i)$ for all sets of ties produced by Judge 2. The following example illustrates the calculation of the tie-adjusted Kendall's tau coefficient.

Example 8.2

Table 8.5 shows the ratings of 6 subjects by Judge1 and Juge2, as well as the associated rankings labeled as Rank1 and Rank2. Judge1 assigned score 8 to both subjects 2 and 3. This led to the common rank 3.5 being assigned to both subjects. This common rank of 3.5 is obtained as the average of 3 and 4, the two ranks that should normally have been assigned to the two subjects. Likewise the score of 1 that Judge 2 assigned to subjects 1, 2, and 4 led to the rank of 2 being assigned to all 3 subjects (i.e. $2 = (1 + 2 + 3)/3$).

Table 8.6 shows the steps for calculating the tie-corrected Kendall's tau. The procedure is very similar to that of Example 8.1 with the exception that any pair of subjects that possesses ties from either judge will be assigned concordance status of 0. This means that the designated pair of subjects will be excluded from the count of concordant and discordant pairs of subjects.

It follows from Table 8.6 that the total counts of concordant and discordant pairs of subjects are respectively given by, $n_{\text{C}} = 7$ and $n_{\text{D}} = 3$. Moreover, the series of ratings from judge 1 has 2 sets of ties {3.5,3.5} and {1.5,1.5} with 2 numbers in each. Consequently, $T_1 = (2^3 - 2) + (2^3 - 2) = 12$. On the other hand, the series of ratings from judge 2 shows a single set of ties {2,2,2} with 3 numbers in it. Therefore, $T_2 = (3^3 - 3) = 24$. Using equation 8.3.2, we can compute the tie-corrected Kendall's tau as follows:

$$\tau^\star = \frac{2 \times (7 - 3)}{\sqrt{6 \times (6-1) - 12} \times \sqrt{6 \times (6-1) - 24}} = 0.7698.$$