Kendall’s Coefficient of Concordance

Kendall’s Coefficient of Concordance, \( W \), is a measure of the agreement between several judges who have rank ordered a set of entities. In this sense, it is similar to an intraclass correlation for ranked data.

As an illustration, we use the same example as for intraclass correlation (from [1]) in which lecturers were asked to mark eight different essays on a percentage scale. Imagine, however, that the lecturers had rank ordered the essays (assigning a rank of 1 to the best and a rank of 8 to the worst) as in Table 1.

If the lecturers agree, then the total ranks for each essay will vary. For example, if all four lecturers thought essay 6 was the best, then the sum of ranks would be only 4, and if they thought essay 3 was the worst, then it would have a total rank of \( 8 \times 4 = 32 \). However, if there is a lot of disagreement then particular essays will have both high and low ranks assigned to them and the resulting totals will be roughly equal. In fact, if there is maximal disagreement between judges, the totals for each essay will be the same.

Kendall’s statistic represents the ratio of the observed variance of the total ranks of the ranked entities to the maximum possible variance of the total ranks. In this example, it would be the variance of total ranks for the essays divided by the maximum possible variance in total ranks of the essays. The first step is, therefore, to calculate the variance of total ranks for essays. To estimate this variance, we use a sum of squared error. In general terms, this is the squared difference between an observation and the mean of all observations:

\[
ss = \sum_{i=1}^{n} (x_i - \bar{x})^2.
\]  

The mean of the total ranks for each essay can be calculated in the usual way, but it is also equivalent to \( k(n+1)/2 \) in which \( k \) is the number of judges and \( n \) is the number of things being ranked.

This gives us:

\[
SS_{\text{Rank Totals}} = (19 - 18)^2 + (15 - 18)^2 + (25 - 18)^2
+ (21 - 18)^2 + (21 - 18)^2 + (14 - 18)^2
+ (17 - 18)^2 + (12 - 18)^2
= 170.
\]  

This value is then divided by an estimate of the total possible variance. We could get this value by simply working out the sum of squared errors for row totals when all raters agree. In this case, the row totals would be 4 (all agree on a rank of 1), 8 (all agree a rank of 2), 12, 16, 20, 24, 28, 32 (all agree a rank of 8). The resulting sum of squares would be:

\[
SS_{\text{Max}} = (4 - 18)^2 + (8 - 18)^2 + (12 - 18)^2
+ (16 - 18)^2 + (20 - 18)^2 + (24 - 18)^2
+ (28 - 18)^2 + (32 - 18)^2
= 672.
\]

The resulting coefficient would be:

\[
W = \frac{SS_{\text{Rank Totals}}}{SS_{\text{Max}}} = \frac{170}{672} = 0.25.
\]

I have done this just to show what the coefficient represents. However, it is a rather cumbersome method because of the need to calculate the maximum possible sums of squares. Thanks to some clever mathematics, we can avoid calculating the maximum possible sums of squares in this long-winded way and simply use the following equation:

\[
W = \frac{12 \times SS_{\text{Rank Totals}}}{k^2(n^3 - n)}.
\]

where \( k \) is the number of judges and \( n \) is the number of things being judged. We would obtain the same answer:

\[
W = \frac{12 \times 170}{16(512 - 8)} = \frac{2040}{8064} = 0.25.
\]

---

Table 1  Eight essays ranked by four lecturers

<table>
<thead>
<tr>
<th>Essay</th>
<th>Dr. Field</th>
<th>Dr. Smith</th>
<th>Dr. Scrote</th>
<th>Dr. Death</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>7</td>
<td>7</td>
<td>3</td>
<td>2</td>
<td>19</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>6</td>
<td>2</td>
<td>1</td>
<td>15</td>
</tr>
<tr>
<td>3</td>
<td>8</td>
<td>8</td>
<td>6</td>
<td>3</td>
<td>25</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>5</td>
<td>7</td>
<td>4</td>
<td>21</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>4</td>
<td>8</td>
<td>5</td>
<td>21</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>6</td>
<td>14</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>2</td>
<td>5</td>
<td>7</td>
<td>17</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>8</td>
<td>12</td>
</tr>
</tbody>
</table>

Mean = 18
A significance test can be carried out using a chi-square statistic with \( n - 1 \) degrees of freedom \( \chi^2_W = k(n - 1)W \). In this case, the test statistic of 7 is not significant.

\( W \) is constrained to lie between 0 (no agreement) and 1 (perfect agreement), but interpretation is difficult because it is unclear how much sense can be made of a statement such as ‘the variance of the total ranks of entities was 25% of the maximum possible’. Significance tests are also relatively meaningless because the levels of agreement usually viewed as good in the social sciences are way above what would be required for significance. However, \( W \) can be converted into the mean **Spearman correlation coefficient** (see [2]):

\[
\tilde{r}_S = \frac{kW - 1}{k - 1}.
\] (7)

If we computed the Spearman correlation coefficient between all pairs of judges, this would be the average value. In this case, we would get \((4 \times 0.25 - 1)/3 = 0\). The clear interpretation here is that on average pairs of judges did not agree!

**References**


**ANDY P. FIELD**