Kendall’s Tau ($\tau_a$ and $\tau_b$)

Another measure of agreement for paired data $(X_i, Y_i)$ is the concordance between pairs of data. Assuming that there are no ties in either $X$ or $Y$, the pairs $(X_i, Y_i)$ and $(X_j, Y_j)$ are concordant if $(X_i - X_j)(Y_i - Y_j) > 0$, and discordant if $(X_i - X_j)(Y_i - Y_j) < 0$. Thus, concordance between pairs of data indicates positive association between the pairs, and discordance indicates negative association.

Kendall’s tau is defined as: $\tau = 2P[(X_i - X_j)(Y_i - Y_j) > 0] - 1$, thus it is a rescaled probability of concordance between variables. It ranges between $-1$ and $1$, with no association corresponding to $\tau = 0$.

**Calculating two types of Kendall’s Tau**

The statistic that is introduced in the text as $r_r$ or $\tilde{r}$ is otherwise known as $\tilde{\tau}_a$. There is a different statistic called $\tilde{\tau}_b$ that is equal to $\tilde{\tau}_a$ when there are no ties in the data, and when there are ties, then $\tilde{\tau}_b > \tilde{\tau}_a$. Calculating by hand is made easier by constructing a table with four columns. The first column has the $X_i$ values sorted in ascending order. The second column has the $Y_i$ values that correspond to the $X_i$ values. The third column calculates the number of $Y_i$ values that are below and larger than the current $Y_i$ value. The fourth column calculates the number of $Y_i$ values that are below and smaller than the current $Y_i$ value. From this completed table, the quantity $C$ is the sum of the third column and $D$ is the sum of the fourth column. Given these quantities, we can calculate $r_r = \tilde{\tau}_a$ as $r_r = \tilde{\tau}_a = 2(C - D)/[n(n - 1)]$. The other Kendall’s Tau statistic, $\tilde{\tau}_b$, is equal to this value when there are no ties. However when the data contains ties then $\tilde{\tau}_b$ is defined as:

$$\tilde{\tau}_b = \frac{2(C - D)}{\sqrt{n^2 - n - \sum s_i(s_i - 1)}\sqrt{n^2 - n - \sum t_i(t_i - 1)},}$$

where $s_i$ is the number of tied $X_i$ values in the $i^{th}$ tie group, and $t_i$ is the number of tied $Y_i$ values in the $i^{th}$ tie group. Since the $s_i$ and $t_i$ values don’t exist if there are no ties, it is easily seen that $\tilde{\tau}_a = \tilde{\tau}_b$ when there are no ties. SAS calculates $\tilde{\tau}_b$, and some prefer it because it can still attain the maximum value of 1 when there are ties. Although the statistics $\tilde{\tau}_a$ and $\tilde{\tau}_b$ may seem less familiar than Spearman’s correlation, some people prefer them to Spearman’s correlation because they estimate a clearly defined population parameter.