PLSC 502: "Statistical Methods for Political Research"

Measures of Association: Ordinal Variables
November 3, 2016

Ordinal Variates: Concordance and Discordance

We'll discuss measures of association for ordinal variables today. For notational purposes, we'll refer to a generic crosstable of two variables Y and X, each of which is assumed to consist of three ordinal categories (coded 1,2, and 3). The respective cell frequencies and the row and column marginals are defined as:

		X			
		1	2	3	
	1	n_{11}	n_{12}	n_{13}	n_{1X}
Y	2	n_{21}	n_{22}	n_{23}	n_{2X}
	3	n_{31}	n_{32}	n_{33}	n_{3X}
		n_{Y1}	n_{Y2}	n_{Y3}	N

The central challenge of measuring association between ordinal variates is how to retain the information present in the ordering of the categories, without giving the numerical values assigned to them cardinal content. We do this through the idea of concordant and discordant pairs of observations in the data. Two observations $i = \{1, 2\}$ in a dataset are said to be *concordant* if:

$$\text{sign}(X_2 - X_1) = \text{sign}(Y_2 - Y_1). \tag{1}$$

Similarly, a *discordant* pair exists where:

$$\text{sign}(X_2 - X_1) = -\text{sign}(Y_2 - Y_1). \tag{2}$$

Thus, for an observation in cell (1,1) of the table above, all observations in cells (2,2), (2,3), (3,2), and (3,3) are concordant with it, because in every instance such observations have both a higher value on Y *and* a higher value of X.

The total number of concordant pairs in the 3×3 table above, then, is equal to

$$N_c = n_{11}(n_{22} + n_{23} + n_{32} + n_{33}) + n_{12}(n_{23} + n_{33}) + n_{21}(n_{32} + n_{33}) + n_{22}(n_{33}). \tag{3}$$

Similarly, the number of discordant pairs is

$$N_d = n_{13}(n_{21} + n_{22} + n_{31} + n_{32}) + n_{12}(n_{21} + n_{31}) + n_{23}(n_{31} + n_{32}) + n_{22}(n_{31}). \tag{4}$$

These numbers will be different in larger tables, but the principle by which they are calculated remains the same. N_c and N_d for the basis for our statistics for association among ordinal variables.

Gamma

Gamma (γ) is the normed difference between the number of concordant and discordant pairs in the data:

$$\gamma = \frac{N_c - N_d}{N_c + N_d} \tag{5}$$

It can also be thought of as the difference between the *proportions* of pairs that are concordant versus discordant:

$$\gamma = \frac{N_c}{N_c + N_d} - \frac{N_d}{N_c + N_d} \tag{6}$$

Note that γ does not count "ties" – no data on pairs are used when (say) $Y_2 > Y_1$ and $X_2 = X_1$, nor when $X_2 = X_1$ and $Y_2 = Y_1$.

- $\gamma \in [-1, 1]$.
- $\gamma = 0 \quad \leftrightarrow$ no association between X and Y, though it can also happen whenever $N_c = N_d$. That is, $\gamma = 0$ is necessary but not sufficient for statistical independence.
- Higher absolute values of γ correspond to stronger associations between X and Y.
- $\gamma = \pm 1.0$ under conditions of (at least) *weak monotonicity* (γ will equal 1.0 whenever, as X increases, Y either increases or stays the same; it will equal -1.0 whenever, as X increases, Y decreases or stays the same.

Inference on γ

The sampling distribution of $\hat{\gamma}$ is Normal, which means that we can use the "usual" approaches to inference, including creation of $(1 - \alpha)$-percent confidence intervals using the normal distribution. We can also test specific hypotheses about the population value of γ by converting our estimate to a z-score:

$$z = (\hat{\gamma} - \gamma)\sqrt{\frac{N_c + N_d}{N(1 - \hat{\gamma}^2)}}. \tag{7}$$

This z-score has a sampling distribution that is $\mathcal{N}(0, 1)$, making inference straightforward.

Kendall's τ

Kendall's τ is similar to γ:

$$\tau = \frac{N_c - N_d}{\frac{1}{2}N(N-1)} \tag{8}$$

You can think of τ as an alternative to γ, one that "norms" the difference between N_c and N_d by a somewhat different number (in this case, the number of all *possible* pairs in the data). As such, the numerator of (8) still signs the statistic, while the denominator scales it.

τ_a, τ_b, and τ_c

τ in (8) is usually called τ_a. Two other variants, τ_b and τ_c, exist in order to "correct" for tied values in the data.

τ_b is generally used for "square" tables; it norms the difference between concordant and discordant pairs to reflect "ties" in the data:

$$\tau_b = \frac{N_c - N_d}{\sqrt{[(N_c + N_d + N_{Y^*})(N_c + N_d + N_{X^*})]}} \tag{9}$$

where N_{Y^*} and N_{X^*} are the number of pairs *not tied* on Y and X, respectively. τ_b is widely used, and is the default in many statistical packages. It's characteristics are:

- $\tau_b \in [-1, 1]$.
- $|\tau_b| = 1.0$ under *strict monotonicity* – that is, if (a) Y increases as X increases (for $\gamma = 1.0$) and (b) there is only one value of Y corresponding to each value of X. Put differently, this means that there are no "ties."
- $\tau_b = 0$ corresponds to no association between X and Y.

τ_c is used for larger, "rectangular" tables.

$$\tau_c = (N_c - N_d) \times \left\{ \frac{2m}{[N^2 2(m-1)]} \right\} \tag{10}$$

where m is the number of rows or columns, whichever is smaller. τ_c is the correct statistic to use when one's table is "rectangular," particularly if one variable has a significantly larger number of categories than the other (e.g., for $2 \times k$ tables where $k \geq 3$).

Because of the way that they norm the differences between N_c and N_d, it will always be the case that

$$\gamma \geq \tau \tag{11}$$

for any of the versions of τ.

The slides contain some examples of the use of γ and τ on some applied data, taken from the September 2008 Big Ten "battleground" poll in Pennsylvania.