Identity of Somers' D and the rank biserial correlation coefficient

Roger Newson

21 February, 2008

1 Formulas

We assume that there is a single sample of N sampling units, partitioned into 2 subsamples (Subsample 1 and Subsample 2), of N_1 and N_2 units, respectively, such that $N_1 + N_2 = N$. For $h \in \{1, 2\}$ and $1 \leq i \leq N_h$, denote by Y_{hi} the outcome measure for the ith unit in Subsample h, and define $X_{hi} = h$, so that the ordinal X–variable indicates membership of the second subsample, rather than the first.

We note that, for each h and i,

$$N = \#\{(j,k) : Y_{jk} < Y_{hi}\} + \#\{(j,k) : Y_{jk} = Y_{hi}\} + \#\{(j,k) : Y_{jk} > Y_{hi}\}, \tag{1}$$

where, for a set S, $\#S$ indicates the cardinality, or number of members, of S. We can now define formally the rank of the ith member of Subsample h as

$$R_{hi} = \frac{1}{2} + \frac{1}{2}\#\{(j,k) : Y_{jk} = Y_{hi}\} + \#\{(j,k) : Y_{jk} < Y_{hi}\}, \tag{2}$$

which implies that ranks can range from 1 to N, and that units in a subset with tied Y–values are assigned the common mean rank that they would have had, if they had been ordered randomly. To simplify the algebra used with mean ranks, we may prefer to work with the linear transformation

$$Q_{hi} = 2R_{hi} - (N+1) = \#\{(j,k) : Y_{jk} < Y_{hi}\} - \#\{(j,k) : Y_{jk} > Y_{hi}\}, \tag{3}$$

as implied by (1) and (2). The sample mean rank, and mean transformed rank, for Subsample h are defined as

$$\bar{R}_h = N_h^{-1} \sum_{i=1}^{N_h} R_{hi}, \quad \bar{Q}_h = N_h^{-1} \sum_{i=1}^{N_h} Q_{hi} = 2\bar{R}_h - (N+1). \tag{4}$$

Note that

$$\begin{aligned}
\bar{Q}_h &= N_h^{-1} \sum_{i=1}^{N_h} \#\{j : Y_{hj} < Y_{hi}\} + N_h^{-1} \sum_{i=1}^{N_h} \#\{j : Y_{2-h+1,j} < Y_{hi}\} \\
&\quad - N_h^{-1} \sum_{i=1}^{N_h} \#\{j : Y_{hj} > Y_{hi}\} - N_h^{-1} \sum_{i=1}^{N_h} \#\{j : Y_{2-h+1,j} > Y_{hi}\} \\
&= N_h^{-1} \sum_{i=1}^{N_h} \left[\#\{j : Y_{2-h+1,j} < Y_{hi}\} - \#\{j : Y_{2-h+1,j} > Y_{hi}\} \right],
\end{aligned} \tag{5}$$

because the terms involving within–sample ordinal contrasts of form $Y_{hi} < Y_{hj}$ and $Y_{hi} > Y_{hj}$ cancel out.

We can now define the rank biserial correlation (RBC) of Cureton (1956) as

$$\text{RBC} = \frac{2}{N}\left(\bar{R}_2 - \bar{R}_1\right) = N^{-1}\left(\bar{Q}_2 - \bar{Q}_1\right). \tag{6}$$

Using (5), we see that

$$\begin{aligned}
\text{RBC} &= \frac{N_1}{N}\frac{1}{N_1 N_2}\sum_{j=1}^{N_2} \#\{k : Y_{1k} < Y_{2j}\} - \frac{N_1}{N}\frac{1}{N_1 N_2}\sum_{j=1}^{N_2} \#\{k : Y_{1k} > Y_{2j}\} \\
&\quad - \frac{N_2}{N}\frac{1}{N_1 N_2}\sum_{k=1}^{N_1} \#\{j : Y_{2j} < Y_{1k}\} + \frac{N_2}{N}\frac{1}{N_1 N_2}\sum_{k=1}^{N_1} \#\{j : Y_{2j} > Y_{1k}\} \\
&= \frac{1}{N_1 N_2}\sum_{j=1}^{N_2}\sum_{k=1}^{N_1} \text{sign}(Y_{2j} - Y_{1k}) \\
&= \hat{D}(Y|X),
\end{aligned} \tag{7}$$

where sign(z) is 1 if $z > 0$, -1 if $z < 0$, and 0 if $z = 0$, and $\hat{D}(Y|X)$ is the sample estimate of Somers' D of Y with respect to X, based on the Y_{hi} and the X_{hi} (Somers, 1962). The sample and population Somers' D parameters are discussed further in Newson (2006) and Newson (2002).

2 References

Cureton EE. Rank–biserial correlation. *Psychometrika* 1956; **21**: 287–290.

Newson R. Confidence intervals for rank statistics: Somers' D and extensions. *The Stata Journal* 2006; **6(3)**: 309–334.

Newson R. Parameters behind "nonparametric" statistics: Kendall's tau, Somers' D and median differences. *The Stata Journal* 2002; **2(1)**: 45–64.

Somers RH. A new asymmetric measure of association for ordinal variables. *American Sociological Review* 1962; **27**: 799-811.