

Association measures in Rapp

2010-03-26, Stig Rosenlund

The parameter `assm` in Proc Taran will generate association measures for all pairs of arguments. Fix one such pair and let ν_{ij} be the sum of exposure for cells where the first argument has values i ($i = 1, \dots, r$) and the second argument has values j ($j = 1, \dots, s$). This notation follows Cramér (1946), section 30.5. Cramér's Table 30.5.1 is

Arguments	1	2	...	s	Total
1	ν_{11}	ν_{12}	...	ν_{1s}	$\nu_{1.}$
2	ν_{21}	ν_{22}	...	ν_{2s}	$\nu_{2.}$
.	—	—	...	—	—
.	—	—	...	—	—
.	—	—	...	—	—
r	ν_{r1}	ν_{r2}	...	ν_{rs}	$\nu_{r.}$
Total	$\nu_{.1}$	$\nu_{.2}$	...	$\nu_{.s}$	n

Here n is total exposure, which is not always an integer. Cramér gives the statistic

$$f^2 = \frac{\chi^2}{n} = \sum_{i,j} \frac{\left(\frac{\nu_{ij}}{n} - \frac{\nu_{i.}\nu_{.j}}{n \cdot n}\right)^2}{\frac{\nu_{i.}\nu_{.j}}{n \cdot n}} = \sum_{i,j} \frac{(\nu_{ij} - \nu_{i.}\nu_{.j}/n)^2}{\nu_{i.}\nu_{.j}}$$

and the inequality, where $q = \min(r, s)$,

$$0 \leq f^2/(q-1) \leq 1$$

The upper limit 1 is attained when and only when each row (when $r \geq s$) or each column (when $r \leq s$) contains one single element different from zero. Then one argument is a function of the other. The lower limit 0 is attained when and only when the arguments are independent of each other in the sense that exposure is multiplicative in the arguments. Thus $f^2/(q-1)$ measures the degree of association between the arguments.

The output from Proc Taran gives the square root (for proper scaling)

$$\text{Assm1} = f/\sqrt{q-1}$$

However, a drawback of Assm1 is that the terms are not weighted with exposure. For some purposes, rows or columns with small exposures might influence the sum too much. Thus a second association measure is also given by Proc Taran, namely

$$\text{Assm2} = \frac{Sq}{(q-1)n} \sum_{i,j} \nu_{ij} \left| \frac{\nu_{ij}}{\nu_{.j}} - \frac{\nu_{i.}}{n} \right| = \frac{Sq}{(q-1)n} \sum_{i,j} \nu_{ij} |\nu_{ij} - \nu_{i.}\nu_{.j}/n|/\nu_{.j}$$

where $S = \max(\text{Assm1})/\max(\text{Assm2}/S)$, with $\max$ taken over all argument pairs. Hence largest Assm1-value = largest Assm2-value. This is the exposure-weighted sum of absolute differences between the proportion of first argument = i when second argument = j and the marginal proportion of first argument = i . As Assm1 it will be 0 when exposure is multiplicative in the arguments.

REFERENCE

CRAMÉR, H. (1946). *Mathematical Methods of Statistics*. Princeton University Press, Princeton.