SOME RELATIONS BETWEEN SIMILARITY COEFFICIENTS: CORRECTION FOR CHANCE SIMILARITY AND K-ADIC FORMULATIONS

A convenient way to summarize the information in two binary (1/0) vectors is the 2 × 2 contingency table given by

$$\begin{array}{cc|c} a & b & p_1 \\ c & d & q_1 \\ \hline p_2 & q_2 & 1 \end{array}$$

where a is the proportion of 1s that both vectors share in the same position, b (c) is the proportion of 1s (0s) in vector 1 (2) and 0s (1s) in vector 2 (1) in the same position, and d is the proportion of 0s that both vectors share in the same position. The information in the above table can be summarized in an index S, called here a coefficient of similarity (resemblance, association). Three well-known examples of S are the simple matching coefficient (or Rand index) given by $S_{\text{SM}} = a + d$ (Sokal and Michener, 1958; Rand, 1971),

$$S_{\text{Dice}} = \frac{2a}{p_1 + p_2} \quad \text{(Dice, 1945)} \quad \text{and Cohen's kappa} \quad S_{\text{Cohen}} = \frac{2(ad - bc)}{p_1 q_2 + p_2 q_1} \quad \text{(Cohen, 1960).}$$

A natural way to correct a coefficient S for the similarity due to chance is given by

$$CS = \frac{S - E(S)}{1 - E(S)}$$

Fleiss (1975) showed that S_{SM} and S_{Dice} (and several other coefficients) become S_{Cohen} when they are corrected for similarity due to chance, given a certain choice of $E(S)$. In the presentation the properties in Fleiss (1975) are considered in a more general context. It is shown which members in a family of coefficients become, after correction for similarity due to chance, either Cohen's kappa or Scott's pi, depending on the expectation used in the correction. In the second part of the presentation formulations of similarity between two or more, say k, binary vectors are presented, that preserve the properties of the first part.

References

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20,* 37-46.

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology, 26,* 297-302.

Fleiss, J. L. (1975). Measuring agreement between two judges on the presence or absence of a trait. *Biometrics, 31,* 651-659.

Rand, W. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association, 66,* 846-850.

Sokal, R. R. and Michener, C. D. (1958). A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin, 38,* 1409-1438.