Agreement between two independent groups of raters

S. Vanbelle and A. Albert

Department of Biostatistics, School of Public Health, University of Liège, Belgium

Although agreement is often searched between two individual raters, there are situations where agreement is needed between two groups of raters. For example, a group of students may be evaluated against a group of experts or two groups of physicians with different specialty may be challenged in diagnosing patients with the same test (positive/negative). Kappa-like agreement indexes are commonly used to quantify agreement between two raters on a nominal or an ordinal scale. They include Cohen's kappa coefficient (Cohen, 1960), the weighted kappa coefficient (Cohen, 1968) and the intraclass kappa coefficient (Kraemer, 1979). To quantify agreement between two groups of raters, the common practice is simply to determine a consensus in each group of raters in order to reduce the problem to the case of two raters. The consensus could be defined by taking the modal category in each group (e.g., van Hoeij and al., 2004) or the median if the scale is ordinal (e.g., Raine and al., 2004). In all cases, however, the question of how to proceed when a consensus can not be reached remains. Moreover, different definitions of consensus may lead to different conclusions (Kraemer and al., 2004) and consensus approaches don't take the variability in the groups into account. We propose a novel agreement index, not requiring any consensus definition and extending the basic concept of Cohen's kappa coefficient to two groups of raters. The agreement index is defined on a population model and the sampling variance is determined by the Jackknife method.

References

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

[2] Cohen J. (1968) *Weighted kappa: nominal scale agreement with provision for scaled disagreement of partial credit*, Psychological bulletin, 70, pp 213-220.

[3] Kraemer H.C. (1979) *Ramifications of a population model for κ as a coefficient of reliability.*, Psychometrika, 44, pp 461-472.

[4] Raine R., Sanderson C., Hutchings A., Carter S., Larking K., Black N. (2004) *An experimental study of determinants of group judgments in clinical guideline development*. Lancet, 364, pp 429-437.

[5] van Hoeij M.J., Haarhuis J.C., Wierstra R.F., van Beukelen P. (2004) *Developing a classification tool based on Bloom's taxonomy to assess the cognitive level of short essay questions*. Journal of Veterinary Medical Education, 31, pp 261-267.

[6] Kraemer H. C., Vyjeyanthi S.P., Noda A. (2004) Agreement Statistics. In R.B. D'Agostino, *Tutorial in Biostatistics vol 1.*, pp. 85-105, John Wiley and Sons.