Supplementary Material

Following the framework and notation given by [1], let us suppose that two raters, i and i'; $i \in \{1,2,\dots,h\}$ each provide a rating for n subjects into one of k mutually exclusive categories indexed by $j \in \{1,2,\dots,k\}$ – for now we assume $k = 2$. This setup can be conceptualized as a 2x2 contingency table, such that $t_{jj'}(ii')$ denotes the number of subjects categorized into category j by rater i and in category j' by rater i'. Subsequently, we have $n = \sum_{j=1}^{k}\sum_{j'=1}^{k} t_{jj'}(ii')$.

Additionally, we have that proportions of the corresponding contingency table with entries

$$u_{jj'}(ii') = \frac{t_{jj'}(ii')}{n} \tag{2}$$

Note that the row and column totals are given by

$$p_{ij} = \sum_{j'}^{k} u_{jj'}(ii') \tag{3}$$

and

$$p_{i'j'} = \sum_{j}^{k} u_{jj'}(ii') \tag{4}$$

respectively.

Cohen's Kappa

Having established the notation, we proceed to define Cohen's Kappa [2], specifically,

$$\hat{\kappa}_{Cohen} = \frac{p_{ii'} - p_e}{1 - p_e} \tag{5}$$

where $p_{ii'}$ is the raw agreement (unadjusted for chance agreement) and given by

$$p_{ii'} = \sum_{j}^{k} u_{jj}(ii') \tag{6}$$

and p_e is the chance agreement which is given by

$$p_e = \sum_{j=1}^{k} p_{ij} p_{i'j} \tag{7}$$

Fleiss' Kappa

Following the notation from above, we assume that there may by an arbitrary number of raters and categories. Subsequently, Fleiss' Kappa, denoted $\hat{\kappa}_{Fleiss}$ is estimated by

$$\hat{\kappa}_{Fleiss} = \frac{\sum_{i=1}^{h-1}\sum_{i'=i+1}^{h}\sum_{j=1}^{k} u_{jj}(ii') - \sum_{i=1}^{h-1}\sum_{i'=i+1}^{h}\sum_{j=1}^{k}\left(\frac{p_{ij}+p_{i'j}}{2}\right)^2}{\frac{h(h-1)}{2} - \sum_{i=1}^{h-1}\sum_{i'=i+1}^{h}\sum_{j=1}^{k}\left(\frac{p_{ij}+p_{i'j}}{2}\right)^2} \tag{8}$$

This was one of the first major attempts in providing a multiple rater statistic for IRA [3].

Light's Kappa and Conger's Kappa

Light presented a IRA statistic for 3 raters [4], which was later generalized to any number of raters by Conger [5]. Here we present the generalization, $\hat{\kappa}_{Conger}$ which is estimated by

$$\hat{\kappa}_{Conger} = \frac{2}{h(h-1)} \sum_{i=1}^{h-1} \sum_{i'=i+1}^{h} \hat{\kappa}_{ii'} \tag{9}$$

where $\kappa_{ii'}$ is Cohen's Kappa between raters i and i'.

Gwet's AC1

Gwet's AC1 statistics provides a estimate for any number of raters and categories [6]. It remains in the form of the Kappa statistics such that

$$\widehat{AC}_1 = \frac{p_a - p_{ey}}{1 - p_{ey}} \tag{10}$$

where p_a, the proportion of agreement is given by

$$P_a = \frac{1}{n} \sum_{s=1}^{n} \sum_{j=1}^{k} \frac{r_{sj}(r_{sj}-1)}{h(h-1)} \tag{11}$$

such that r_{sj} is the number of raters who classified the sth subject into the jth category. Additionally,

$$p_{ey} = \frac{1}{k} \sum_{j=1}^{k} \pi_j (1 - \pi_j) \tag{12}$$

where

$$\pi_j = \frac{1}{n} \sum_{i=1}^{n} \frac{r_{sj}}{h} \tag{13}$$

Final Remarks

We have intentionally omitted any formalisms around the variance since there is a great deal of heterogeneity on the topic, and any treatment thereof here would necessarily be an incomplete one. That is, several estimators of the variance have been proposed for each of the above statistics. The following references, may however, be of interest as a starting point [7–10].

References

1. Warrens, M.J. Inequalities between multi-rater kappas. *Adv. Data Anal. Classif.* **2010**, *4*, 271–286, doi:10.1007/s11634-010-0073-4.
2. Cohen, J. A Coefficient of Agreement for Nominal Scales. *Educ. Psychol. Meas.* **1960**, *20*,

37–46, doi:10.1177/001316446002000104.
3. Fleiss, J. Measuring Nominal Scale agreement amongst many raters. *Psychol. Bull.* **1971**, *76*, 378–382.
4. Light, R.J. Measures of response agreement for qualitative data: Some generalizations and alternatives. *Psychol. Bull.* **1971**, *76*, 365–377, doi:10.1037/h0031643.
5. Conger, A.J. Integration and generalization of kappas for multiple raters. *Psychol. Bull.* **1980**, *88*, 322–328, doi:10.1037/0033-2909.88.2.322.
6. Gwet, K.L. *Handbook of Inter-Rater Reliability*; 4th ed.; Advanced Analytics, 2014; ISBN 978-0-9708062-8-4.
7. Banerjee, M.; Capozzoli, M.; Mcsweeney, L.; Sinha, D. Beyond kappa : A rev interrater agreemen. **2019**, *27*, 3–23.
8. Blood, E.; Spratt, K.F. Disagreement on Agreement : Two Alternative Agreement Coefficients. *SAS Glob. Forum 2007* **2007**, 1–12, doi:10.3109/00016922209137206.
9. Grassano, L.; Pagana, G.; Daperno, M.; Bibbona, E.; Gasparini, M. Asymptotic distributions of kappa statistics and their differences with many raters, many rating categories and two conditions. *Biometrical J.* **2018**, *60*, 146–154, doi:10.1002/bimj.201700016.
10. Gwet, K.L. Computing inter-rater reliability and its variance in the presence of high agreement. *Br. J. Math. Stat. Psychol.* **2008**, *61*, 29–48, doi:10.1348/000711006X126600.