Agreement between 2 independent groups of raters

S. Vanbelle
Biostatistics, University of Liège, Belgium

Summary

A novel agreement index, based on a population model is proposed. It extends the basic concept of Cohen's kappa coefficient [1] for two groups of raters and reduces to it in case of one rater in each group. Sampling variability is derived by the Jackknife method [2].

Example: Script Concordance Test (SCT) [3]

Aim: Evaluate the ability of students to solve unclear clinical situations

Test: N items (description of a situation + diagnosis assumption)

Principle: Evaluate impact of new information on the assumption using a 5-point Likert scale: (-2) Eliminated $\rightarrow$ (+2) Only possibility

Particularity: No correct answers $\rightarrow$ experts define a "gold standard"

Position of the problem	Study [4]
Group G_1 of R_1 raters	Medical experts ($R_1 = 10$)
Group G_2 of R_2 raters	Students in medicine
N items	SCT with 48 items
K-categorical scale	5-point Likert scale
Agreement(G_1,G_2)=?	Agreement(experts,students)=?

Study aim: Year 7 students ($R_2 = 27$) better than year 5 ($R_2 = 20$)?

Case of 2 isolated raters (Cohen's κ coefficient [1])

2 raters, N items, 1 K-category scale $\rightarrow$ $K \times K$ contingency table

	Rater 2		
Rater 1	Yes	No	Total
Yes	0.55*	0.12	0.67
No	0.11	0.22	0.33
Total	0.66	0.34	1

*$p_{jk} = n_{jk}/N, j, k = 1, \cdots, K$

- $p_o = 0.55 + 0.22 = 0.77$
  The 2 raters agree on 77% of the items.
- $p_e = 0.66 \times 0.67 + 0.34 \times 0.33 = 0.55$
  55% of agreements only expected by chance.
- $\hat{\kappa} = (0.77 - 0.55)/(1 - 0.55) = 0.69$

Observed proportion of agreement: $p_o = \sum_{j=1}^{K} p_{jj}$

Proportion of agreement expected by chance: $p_e = \sum_{j=1}^{K} p_{j.}p_{.j}$

Cohen's κ coefficient (agreement corrected for chance): $\hat{\kappa} = \frac{p_o - p_e}{1 - p_e}$

Interpretation: $\hat{\kappa} = 1$ Perfect agreement
$\hat{\kappa} = 0$ Agreement not better than chance
$\hat{\kappa} < 0$ Agreement lower than chance.

Existing methods for 2 groups of raters

- **Consensus**: Summarize responses of each group in 1 quantity
  $\rightarrow$ Agreement between the 2 consensuses

Example:

Rule	Consensus category
Majority	category chosen by the majority of the raters in the group
x%	category chosen by at least x% of the raters in the group

Drawbacks: (Orange cell = Consensus category)

	Majority rule		80% rule	
Category of a binary scale	No	Yes	No	Yes
Not always defined	50[a]	50	60	40
Different rules $\rightarrow$ different conclusions	70	30	70	30
Different consensus strength $\rightarrow$ same answer	60	40	80	20
(variability in the group not taken into account)	90	10	90	10

[a]% of raters selecting the category

- **Schouten** [5]: Consider all pairs of raters with 1 rater of each group

Principle: $\bar{p}_o$ = mean p_o between all pairs
$\bar{p}_e$ = mean p_e between all pairs
$\hat{\kappa} = (\bar{p}_o - \bar{p}_e)/(1 - \bar{p}_e)$

Drawbacks:

Many pairs
Definition of perfect agreement too restrictive (see later)

New agreement index (Binary scale)

Population $\mathcal{R}_g$ ($g = 1, 2$) of raters and $\mathcal{I}$ of items

- $X_{ir,g} = 1$ if rater $r \in \mathcal{R}_g$ classifies item i in category 1
- $P(X_{ir,g} = 1) = E(X_{ir,g}|\mathcal{I}) = P_i, g$ over $\mathcal{R}_g$ and $E(P_{i,g}) = \pi_g$, $var(P_{i,g}) = \sigma_g^2$ over $\mathcal{I}$
- In $\mathcal{R}_g$, $ICC_g = \sigma_g^2/\pi_g(1 - \pi_g)$ [6] (=1 if perfect agreement in $\mathcal{R}_g$)

Theoretical agreement: $\Pi_T = E[P_{i,1}P_{i,2} + (1 - P_{i,1})(1 - P_{i,2})]$

Agreement expected by chance: $\Pi_E = \pi_1\pi_2 + (1 - \pi_1)(1 - \pi_2)$

Perfect agreement when $P_{i,1} = P_{i,2} = P_i$ with $E(P_i) = \pi$, $var(P_i) = \sigma^2$
$\rightarrow \Pi_T = \Pi_M = 1 - 2\pi(1 - \pi)(1 - ICC)$

New agreement index: $\kappa = (\Pi_T - \Pi_E)/(\Pi_M - \Pi_E)$

Comparison of the methods

Index	Perfect agreement	Π_M
New	Same probability distribution in both populations	≤ 1
Schouten	+ Perfect agreement in both populations	$= 1$
Consensus	+ consensus always possible	$= 1$

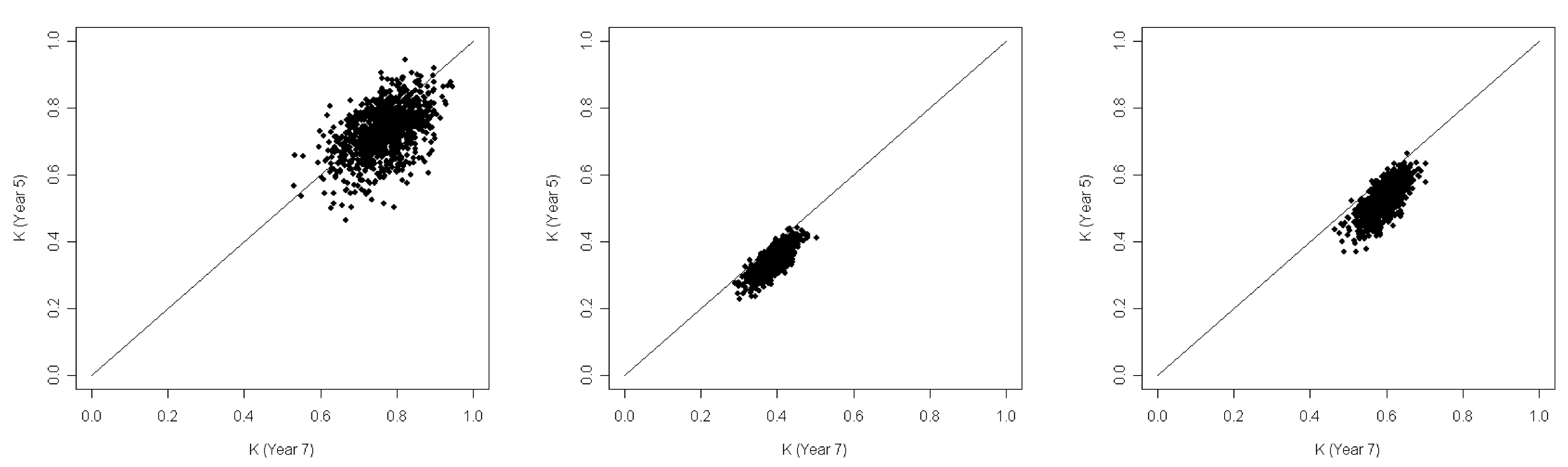
$\rightarrow$ Schouten's index = special case when $ICC = 1$.

Results of the SCT example ($\hat{\kappa} \pm SE$)

Method	N	Year 5	Year 7	p-value[a]
Majority rule	39	0.73 ± 0.08	0.77 ± 0.07	0.58
Schouten index	48	0.35 ± 0.04	0.40 ± 0.03	0.028
New index	48	0.53 ± 0.05	0.60 ± 0.04	0.030

[a]Comparison with the bootstrap method (1000 iterations) [7]

Consensus | Schouten | New index



Conclusion: Year 7 students better agree with experts than Year 5.

Discussion

- New index quantifies the agreement between 2 groups of raters
- Based on a population model
- Weighted and intraclass versions were derived
- Possess same interpretation and properties as Cohen's κ
- Reduce to Cohen's kappa coefficient when 1 rater in each group
- Better than consensus approach (always defined and account for variability in the groups)
- Schouten's index is a special case (more restrictive definition of perfect agreement)
- Better estimate Π_M?

Bibliography

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

[2]Efron, B. & Tibshirani, R.J. (1993). An introduction to the bootstrap. *Chapman and Hall, New York.*

[3]Charlin, B., Gagnon, R., Sibert, L., & Van der Vleuten, C. (2002). Le test de concordance de script: un instrument d'évaluation du raisonnement clinique. *Pédagogie Médicale*, *3*, 135–144.

[4]Vanbelle, S., Massart, V., Giet, G., & Albert, A. Test de concordance de script: un nouveau mode d'établissement des scores limitant l'effet du hasard. *Pédagogie Médicale*, *8*, 71–81.

[5]Schouten H.J.A. (1982). Measuring pairwise interobserver agreement when all subjects are judged by the same observers. *Statistica Neerlandica*, *36*, 45–61.

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

[7] Vanbelle, S. & Albert, A. A bootstrap method for comparing correlated kappa coefficients . *Journal of Statistical Computation and Simulation*, in press.

sophie.vanbelle@ulg.ac.be