A Model-based Approach to Assessing Inter-rater Agreement

Rui Wu, supervised by Damjan Vukcevic

THE UNIVERSITY OF MELBOURNE

Introduction

In some research fields, especially medical research, data-collected are usually categorical. Using this type of data requires confidence in the agreement between data-collectors. Hence, Cohen's Kappa appeared as a tool for assessing this agreement. However, in recent years, academics realise there are several constraints of Cohen's Kappa[1], which limits its application. Hence, in this project we explored another way of assessing the inter-rater agreement based on the Dawid-Skene model[2].

Methodology

Dawid-Skene Model Parameters

π_k: The prevalence of category in the sampled population

$\theta_{j,k,k'}$: The probability that rater j rates item with true class k as k'

z_i: The true class of item i (include likelihood function or not)

Cohen's Kappa $\boldsymbol{\kappa}$

A popular way of assessing agreement between raters

p_0: observed agreement between raters

p_e: estimated chance agreement between raters assuming they are independent

$$\boldsymbol{\kappa} = \frac{p_0 - p_e}{1 - p_e}$$

Rater accuracy

$$\Pr(\text{rater } j \text{ rates correctly for item } i) = \sum_{k=1}^{K} \theta_{j,k,k} \cdot \pi_k$$

Inter-rater agreement

$$\boldsymbol{A} = \Pr(\text{raters } j \text{ and } j' \text{ rate the same for item } i)$$

$$= \sum_{k'=1}^{K} \sum_{k=1}^{K} \theta_{j,k,k'} \cdot \theta_{j',k,k'} \cdot \pi_k$$

$$A_{chance} = \Pr(\text{raters } j \text{ and } j' \text{rate the same by chance})$$

$$= \sum_{k'=1}^{K} \sum_{k=1}^{K} \sum_{k''=1}^{K} \theta_{j,k,k'} \cdot \theta_{j',k'',k'} \cdot \pi_k \cdot \pi_{k''}$$

$$\kappa' = \frac{A - A_{chance}}{1 - A_{chance}}$$

1. The rater package[4] was used to fit the Dawid-Skene model[2] on data sets to obtain estimates of $\theta_{j,k,k'}$ and π_k
2. Inter-rater agreement was calculated using $\boldsymbol{\kappa}$, **A** and $\boldsymbol{\kappa}$ '
3. Rater accuracy and other values were calculated to help with investigation

Data Sets

- "Anesthesia" was obtained from the original paper of Dawid-Skene model[2]
- The "simulated" data were generated using estimates of $\theta_{1,k,k'}$, $\theta_{2,k,k'}$ and π from the Dawid-Skene model fitted to the "Anesthesia" data. It comprises of 1000 simulated ratings of rater1 and rater2

Results & Discussion

Comparing κ and **A** as tools for assessing inter-rater agreement

	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5
Rater 1	1.00	0.41	0.46	0.55	0.47
Rater 2	0.41	1.00	0.48	0.58	0.48
Rater 3	0.46	0.48	1.00	0.53	0.59
Rater 4	0.55	0.58	0.53	1.00	0.56
Rater 5	0.47	0.48	0.59	0.56	1.00

Table 1. $\boldsymbol{\kappa}$ Matrix with data "Anesthesia"

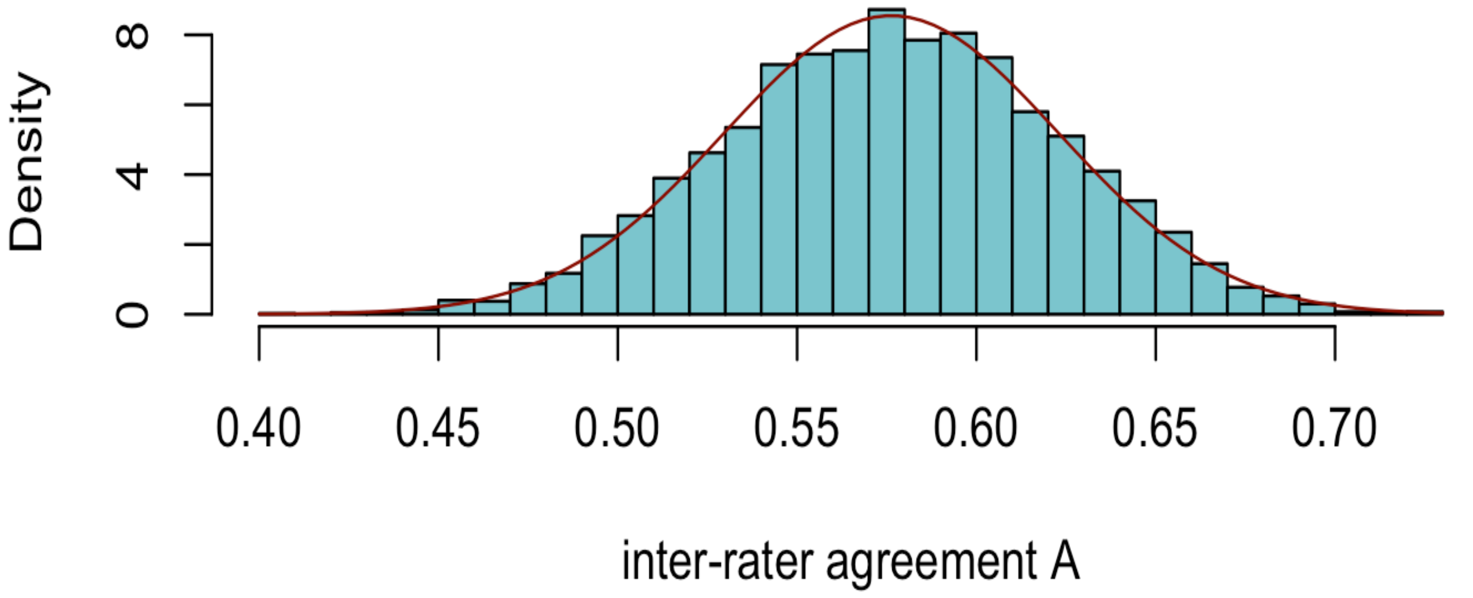
	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5
Rater 1	0.71	0.57	0.65	0.65	0.63
Rater 2	0.57	0.50	0.53	0.53	0.52
Rater 3	0.65	0.53	0.60	0.59	0.59
Rater 4	0.65	0.53	0.59	0.59	0.58
Rater 5	0.63	0.52	0.58	0.58	0.57

Table 2. Inter-rater agreement ***A*** matrix with data "Anesthesia", calculated using the point estimates of θ and π

$\boldsymbol{\kappa}$ heavily relies on the reduction of marginal sums, which are considered the estimates of chance agreement[3]. This makes it hard to interpret and sometimes over conservative.

Thus, estimating rater agreement using A is more reliable and sensible in this situation.

Graph 1. Posterior distribution of A between rater1 and rater 2



Agreement between rater 1 and rater 2	$\boldsymbol{\kappa}$	$\boldsymbol{A}$	$\boldsymbol{\kappa}$
Anesthesia	0.41	0.57	0.30
Simulated	0.41	0.59	0.30

The table above compares κ , **A**, κ' of the same rater pair in original "Anesthesia" and the simulated Anesthesia data sets. κ and κ' are the same for both data sets and A varies slightly when the sample is larger. Hence, both statistics perform rigorously even when the sample is small.

Both $\boldsymbol{\kappa}$ and **A** show that raters 1 and 2 have a medium level of agreement.

Individual Accuracy of Raters		
	Rater1	Rater2
Percentage accuracy	0.8220	0.6450
By rater accuracy formula	0.7531	0.7117

Both rater accuracies are decent. In Table 3, this pair of raters are shown to having a greater level of disagreement when items assessed are of less prevalence

	Rate 1	Rate 2	Rate 3	Rate 4
#Agreement over #disagreement	1.47	0.88	0.50	0.47
prevalence	0.38	0.41	0.14	0.08
$\theta_{1,k,k}$	0.86	0.85	0.79	0.69
$\theta_{2,k,k}$	0.75	0.58	0.63	0.65

Table 3. Disagreement ratios of rater 1 and 2 when giving a specific rating

Remarks

- Both $\boldsymbol{\kappa}$, A and $\boldsymbol{\kappa}'$ perform rigorously regardless of sample size
- $\boldsymbol{\kappa}$ removes marginal sums, which has resulted in several issues
  - Is marginal sums representative of chance agreement
  - Chance agreement made by raters are acceptable as the main concern is about them making the same and correct ratings

  Hence, we suggest A as a more comprehensive statistic for assessing inter-rater agreement
- Inter-rater agreement values can be noisy for
  - Contradicting rater accuracy figure
  - Their tendency to underestimate the agreement when rare categories present

References

1. Cohen, J. (1960), A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement, 20(1), 37–46.
2. Dawid, A.P. and Skene, A.M. (1979), Maximum Likelihood Estimation of Observer Error-Rates Using the *EM* Algorithm. Journal of the Royal Statistical Society: Series C (Applied Statistics), 28: 20-28.
3. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb).2012;22(3):276-82. PMID: 23092060; PMCID: PMC3900052.
4. Pullin, J. et al (2020), Statistical Model of Repeated Categorical Rating: The R Package Rater. arXiv: 2010.09335.