An Extension of Cohen's Kappa for Clustered Data and Group Sequential Testing

Mary M. Ryan & Daniel L. Gillen

Department of Statistics, University of California, Irvine

Motivation

- **SPOT GRADE scale** developed to standardize severity of blood loss in wounds[1]
  - Surface bleed severity scale (SBSS) with 6 categories, 0-5
  - Scores defined by ranges of flux/flow rate of blood from wound (0="hemostasis"; 5="gushing")
- 14 surgeons used to validate scale
  - 36 training videos
  - 36 testing videos, each viewed 3 times; **repeated measures induced**
- **Kappa statistic**[2] used to assess inter- and intra-rater reliability
- Varying flux and wound size in videos of same category may **introduce heterogeneity** to within-category classification probabilities

Cohen's Kappa

- $\kappa = \frac{p_o - p_e}{1 - p_e} \in (-1, 1)$
  - $p_o = \sum_{i=1}^{k} p_{ii}$
  - $p_e = \sum_{i=1}^{k} p_{i\cdot} p_{\cdot i}$
- Used to assess above-likelihood-chance of two raters agreeing
- By CLT[3]:
  $\sqrt{n}(\kappa - \kappa_0) \sim N\left(0, \sigma^2_\kappa\right)$,
- To map onto $\mathbb{R}$:
  $f(\kappa) = ln\left(\frac{1+\kappa}{1-\kappa}\right) \equiv \varphi$

Operating Characteristics of Kappa under Heterogeneous Samples

- 1,000 simulations; 50 surgeons
- 4 with-category heterogeneity settings: none, low, medium, high
- 6 videos per category; each video viewed 3 times

Video Heterogeneity	$\kappa = 0.8$	
	Variance Ratio	Coverage
None	1.067	0.958
Low	1.228	0.968
Medium	1.494	0.984
High	2.036	0.995

Table 1: Analytic/empirical variance ratios and coverage probabilities under $\kappa = 0.8$ and different video heterogeneity settings when each video was viewed three times.

- Increases of video heterogeneity, combined with data clustering, **inflates analytic variance estimate** relative to true variance, leading to conservative inference
- We proposed a bootstrap variance estimator, $\hat{\sigma}^2_B$ (Algorithm 1), to correct this (results shown in Table 2)

Algorithm 1:
for *b in B* **do**
  Randomly choose n surgeons, with replacement;
  Add together all sampled surgeon contingency tables;
  Find statistic, κ_b;
end
Calculate $\hat{\sigma}^2_B = var(\vec{\kappa})$

Video Heterogeneity	$\kappa = 0.8$	
	Variance Ratio	Coverage
None	0.965	0.938
Low	0.962	0.937
Medium	0.983	0.939
High	0.994	0.941

Table 2: Theoretical/empirical variance ratios and coverage probabilities under $\kappa = 0.8$ and different video variability settings, when 200 bootstrap samples were performed at each simulation and each video was viewed three times.

Heterogeneity in SPOT GRADE

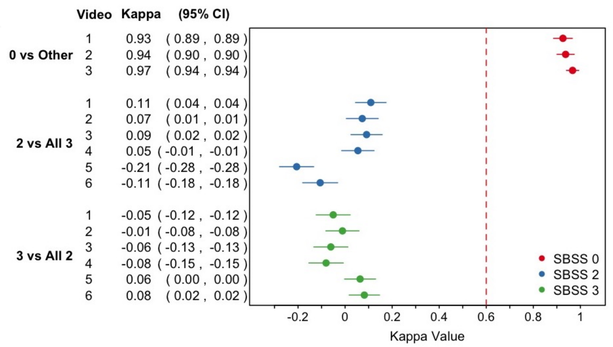

Figure 1: Surgeons ability to correctly identify the SPOT GRADE category for specific videos of category 0 vs observations from all other categories (top), category 2 vs all other category 3 videos (middle), and category 3 vs all other category 2 videos (bottom).

Kappa using Group Sequential Design

- Wanted to explore how transformed Kappa statistic performed in various group sequential design scenarios
- $\hat{\sigma}^2_B$ was used for inference

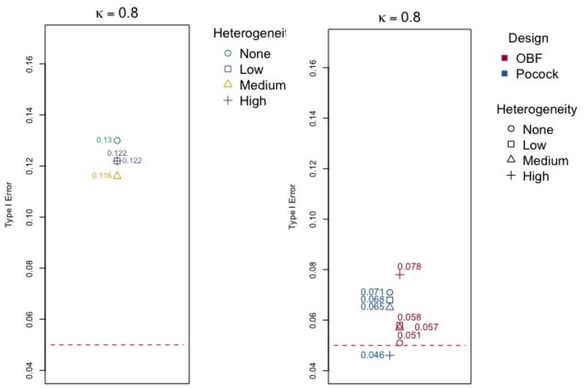

Figure 2: Type I error rates for bootstrapped Kappa in (a) naive, multiple testing and (b) group sequential settings. Under bootstrap, 200 bootstrap samples were performed. Six videos were rated per category, with each video being viewed three times.

Conclusions

- Data clustering, paired with heterogeneity of classification probabilities within a category, inflates Kappa's analytic variance estimate (Table 1)
  - Bias can be corrected with bootstrap variance estimator (Table 2)
- It is **likely that homogeneous classification probabilities within each category is an unrealistic assumption** (Figure 1)
  - Large κ among SBSS 0 videos (high rate of correct classification), with little variation
  - Low κ among SBSS 2 and 3 videos (low rate of correct classification), with higher amounts of variation (some videos easier to classify than others)
- Rater agreement studies can be conducted using group sequential designs
  - Significant gains in study efficiency stand to be made (Figure 2)

References

[1] Spotnitz WD et al. The SPOT GRADE: A new method for reproducibly quantifying surgical wound bleeding. Spine 2018; 43(11):E664.

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

[3] Fleiss JL et al. Large sample standard errors of kappa and weighted kappa. Psychological Bulletin 1969; 72(5):323-327.

Acknowledgements

This work was funded by NIA AG000096. We would like to thank investigators at Biom'up, SA, for the use of their study data.