Optimal Data Analysis
Vol. 5 (October 28, 2016), 160-161

2155-0182/10/$3.00

ODA *vs*. π and κ: Paradoxes of Kappa

Paul R. Yarnold, Ph.D.
Optimal Data Analysis, LLC

Widely-used indexes of inter-rater or inter-method agreement, π and κ sometimes produce unexpected results called the paradoxes of kappa. For example, prior research obtained four legacy agreement statistics (κ, Scott's π, G-index, Fleiss's generalized π) for a 2x2 table in which two independent raters failed to jointly classify any observations into the "negative" rating-class category: two indexes reported $\geq$ 88.8% overall agreement and the other two reported $\leq$ -2.3% overall agreement.[1] ODA sheds new light on this paradox by testing confirmatory and exploratory hypotheses for these data, separately modeling the ratings made by each rater, and separately maximizing model *predictive accuracy* normed for chance (ESS; 0=inter-rater agreement expected by chance, 100=perfect agreement) as well as model *overall accuracy* that is not normed for chance (PAC; 0=no inter-rater agreement, 100=perfect agreement).[2-7]

Consistent with many legacy multivariable methods used today, early research on explicitly optimal methods sought to maximize the overall *p*ercentage *a*ccurate *c*lassification (PAC) of a statistical model: if zero observations in the sample are correctly classified (predicted) then PAC=0, and if all sample observations are correctly classified then PAC=100. In contrast to the ESS index of model predictive accuracy, PAC is *not* normed against chance.[2-9] Analyses presented herein address the data in Table 1.

Table 1: Pathological Data Example[1]

	Rater B	
Rater A	Negative	Positive
Negative	0	2
Positive	5	118

The *confirmatory* alternative hypothesis is that raters' ratings agree, the null hypothesis is that the raters' ratings are unrelated.[2] In the first pair of analyses Rater A's ratings (negative, positive) was the class variable, and B's ratings (negative, positive) was the categorical attribute. The model that maximized ESS was: if B's rating=negative, predict A's rating=negative; otherwise predict A's rating=positive. Model sensitivities were 0% (0/2) for negative ratings, and 95.9% (118/123) for positive ratings: thus ESS= -4.07 (worse than expected by chance), p<0.99.

The identical model also maximized overall PAC of 94.4% (118/125), p<0.99.

The next pair of confirmatory analyses used B's ratings as class variable, and A's ratings as categorical attribute. The model that maximized ESS was: if A's rating=negative, predict B's rating=negative; otherwise predict

2155-0182/10/$3.00

B's rating=positive. Model sensitivities were 0% (0/5) for negative ratings, and 98.3% (118/120) for positive ratings: ESS= -1.67, $p<0.99$.

The same model also maximized overall PAC of 94.4% (118/125), $p<0.99$.

The *exploratory* alternative hypothesis is raters' ratings are related, the null hypothesis is raters' ratings are unrelated.[2] The first pair of analyses used A's ratings as class variable, and B's ratings as categorical attribute. The model maximizing ESS was: if B's rating=negative, predict A's rating=positive; otherwise predict A's rating=negative. Model sensitivities were 100% (2/2) for negative ratings, and 4.1% (5/123) for positive ratings: ESS=4.07, $p<0.99$.

In contrast, the model maximizing PAC was: if A's rating=negative, predict B's rating=negative; otherwise predict B's rating=positive. Model sensitivities were 0% (0/2) for negative ratings, and 95.9% (118/123) for positive ratings: PAC=94.4%, $p<0.99$.

The final pair of exploratory analyses used B's ratings as class variable, and A's as categorical attribute. The model maximizing ESS was: if A's rating=negative, predict B's rating=positive; otherwise predict B's rating=negative. Model sensitivities were 100% (5/5) for negative ratings, and 1.7% (2/120) for positive ratings: ESS=1.67, $p<0.99$.

In contrast, the model maximizing PAC was: if A's rating=negative, predict B's rating=negative; otherwise predict B's rating=positive. Model sensitivities were 0% (0/5) for negative ratings, and 98.3% (118/120) for positive ratings: PAC=94.4%, $p<0.99$.

No ODA model considered was statistically reliable, so it is concluded that the raters' ratings did not agree at a level exceeding what is expected by chance in this application.

References

[1]Gwet KI (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology, 61*, 29–48.

[2]Yarnold PR, Soltysik RC (2016). *Maximizing predictive accuracy*. Chicago, IL: ODA Books. DOI: 10.13140/RG.2.1.1368.3286

[3]Yarnold PR, Soltysik RC, McCormick WC, Burns R, Lin EHB, Bush T, Martin GJ (1995). Application of multivariable optimal discriminant analysis in general internal medicine. *Journal of General Internal Medicine, 10*, 601-606.

[4]Yarnold PR, Soltysik RC, Lefevre F, Martin GJ (1998). Predicting in-hospital mortality of patients receiving cardiopulmonary resuscitation: Unit-weighted MultiODA for binary data. *Statistics in Medicine, 17*, 2405-2414.

[5]Soltysik RC, Yarnold PR (2010). Two-group MultiODA: Mixed-integer linear programming solution with bounded *M*. *Optimal Data Analysis*, *1*, 31-37.

[6]Yarnold PR (2016). Identifying the descendant family of HO-CTA models by using the minimum denominator selection algorithm: Maximizing ESS versus PAC. *Optimal Data Analysis*, *5*, 53-57.

[7]Yarnold PR (2016). Pruning CTA models to maximize PAC. *Optimal Data Analysis, 5*, 58-61.

[8]Grimm LG, Yarnold PR (1995). *Reading and understanding multivariate statistics*. Washington, DC: APA Books.

[9]Grimm LG, Yarnold PR (1995). *Reading and understanding more multivariate statistics*. Washington, DC: APA Books.

Author Notes

Publically-available data were analyzed. No conflict of interest was reported.

Mail: Optimal Data Analysis, LLC
6348 N. Milwaukee Ave., #163
Chicago, IL 60646