Optimal Data Analysis
Vol. 3 (November 23, 2014), 121-123

2155-0182/10/$3.00

UniODA *vs.* Kendall's Coefficient of Concordance (*W*): Multiple Rankings of Multiple Movies

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

Kendall's coefficient of concordance *W* is a non-parametric statistic used to assess agreement in rankings of multiple stimuli made by multiple raters. A normalization of the test statistic for Friedman's non-parametric alternative to ANOVA with repeated measures, *W* ranges from 0 (no agreement) to 1 (complete agreement).[1-3] Problems with *W* include complications arising from tied rankings, invalid Type I error estimates for small samples, the absence of an effect strength index normed versus chance, and the absence of an explicit model relating ranks to movies. The use of UniODA to assess inter-rater agreement is demonstrated as an alternative approach that overcomes these issues.

W is typically used in applications with three or more raters, while Cohen's Kappa is typically used to assess agreement for a pair of ratings.[4-6] *W* is not a correlation coefficient but it is linearly related to the mean Spearman's rank correlation coefficient obtained for all pairs of rankings. When computing *W* tied rankings are replaced by the mean of the ranks that would have be assigned if no ties had occurred, serving to reduce the magnitude of *W*. When there are many ties (this is a qualitative assessment) a correction is employed. Statistical significance of *W* is commonly assessed using chi-square, but when this methodology is invalid because the minimum expectation is small, bootstrap assessment is instead appropriate.[2,7]

Demonstrated in this note, UniODA provides an exact alternative methodology for assessing omnibus inter-rater agreement in applications involving ten or fewer raters: tied rankings are not an issue; exact Type I error is provided regardless of the sample size; a measure of effect strength normed versus chance is given; and an explicit model discriminating stimuli on the basis of rater rankings is presented.[6,8]

Data involved rankings of eight movies made by seven raters.[3] For this example W = 0.635 (mean Spearman r=0.574), suggesting moderate levels of inter-rater agreement. Chi-square indicated statistically marginal rejection of the null hypothesis of no agreement among raters: p<0.0591. However, because the

2155-0182/10/$3.00

expected values for ratings of movies C (4.6), G (1.7) and H (1.9) were all less than five, chi-square is an invalid approximation in this application.[7]

These data were analyzed using UniODA[6,9] and the following maximum-accuracy model was identified:

```
IF RANK <= 1.5 THEN MOVIE = H
IF 1.5 < RANK <= 2.25 THEN MOVIE = G
IF 2.25 < RANK <= 2.75 THEN MOVIE = D
IF 2.75 < RANK <= 3.5 THEN MOVIE = F
IF 3.5 < RANK <= 4.25 THEN MOVIE = C
IF 4.25 < RANK <= 5.25 THEN MOVIE = E
IF 5.25 < RANK <= 7.25 THEN MOVIE = B
IF 7.25 < RANK THEN MOVIE = A
```

Normed performance indices[6,10] revealed that this model yielded moderate ESS of 34.7%, and moderate ESP of 36.6%. This performance resulted in statistically significant rejection of the null hypothesis of no agreement among raters: $p<0.024$ (300 Monte Carlo experiments indicated 98.4% confidence for exact $p<0.05$).

In contrast to the analysis using *W*, for UniODA the Type I error rate was valid (the use of bootstrap methodology to estimate *p* for *W* would remedy this issue); a normed index of effect strength was provided; and a model explicitly identifying how ranks discriminated movies was identified.

Not illustrated for lack of requisite data, using UniODA in this application a weighted analysis may also be conducted.[6] For example, if raters individually provided an estimate of the certainty of each of their rankings using a continuous index (e.g., a Likert-type rating, or a percentage score), then a maximum-accuracy measure of their certainty-adjusted concordance would be produced. In applications involving more than ten raters, or in which more detailed analysis of concordance between pairs of raters is desired, then the UniODA-based alternative to Kappa may be employed.[11,12]

References

[1]Kendall MG, Babington SB (1939). The problem of *m* rankings. *The Annals of Mathematical Statistics*, 10, 275–287.

[2]Legendre P (2005). Species associations: The Kendall Coefficient of Concordance revisited. *Journal of Agricultural, Biological and Environmental Statistics*, 10, 226–245.

[3]http://www.real-statistics.com/reliability/kendalls-w/

[4]Yarnold PR (2014). UniODA *vs*. weighted kappa: Evaluating concordance of clinician and patient ratings of the patient's physical and mental health functioning. *Optimal Data Analysis*, *3*, 12-13.

[5]Yarnold PR (2014). UniODA *vs*. kappa: Evaluating the long-term (27-year) test-retest reliability of the Type A Behavior Pattern. *Optimal Data Analysis*, *3*, 14-16.

[6]Yarnold PR, Soltysik RC (2005). *Optimal data analysis: A guidebook with software for Windows*, Washington, DC, APA Books.

[7]Yarnold JK (1970). The minimum expectation in X^2 goodness of fit tests and the accuracy of approximations for the null distribution. *Journal of the American Statistical Association*, *65*, 864-886.

[8]Yarnold PR, Soltysik RC (2010). Optimal data analysis: A general statistical analysis paradigm. *Optimal Data Analysis*, *1*, 10-22.

[9]Control commands[6] used to conduct analysis:

```
VARS MOVIE RANK;
CLASS MOVIE;
ATTR RANK;
MCARLO ITER 1000 TARGET .05
STOP 95.0 STOPUP 95.0;
GO;
```

2155-0182/10/$3.00

[10]Yarnold PR (2013). Standards for reporting UniODA findings expanded to include ESP and all possible aggregated confusion tables. *Optimal Data Analysis*, *2*, 106-119.

[11]Yarnold PR (2014). How to assess inter-observer reliability of ratings made on ordinal scales: Evaluating and comparing the Emergency Severity Index (Version 3) and Canadian Triage Acuity Scale. *Optimal Data Analysis*, *3*, 42-49.

[12]Yarnold PR (2014). How to assess the inter-method (parallel-forms) reliability of ratings made on ordinal scales: Evaluating and comparing the Emergency Severity Index (Version 3) and Canadian Triage Acuity Scale. *Optimal Data Analysis*, *3*, 50-54.

Author Notes

E-mail: Journal@OptimalDataAnalysis.com.

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