Optimal Data Analysis
Vol. 8 (May 18, 2019), 103-106

2155-0182/10/$3.00

Fixed *vs*. Relative Optimal Discriminant Thresholds: Pairwise Comparisons of Raters' Ratings for a Sample

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

Foundational to the ODA algorithm when used with an ordered attribute is the identification of the optimal threshold—the specific cutpoint that yields the most accurate (weighted) classification solution for a sample of observations.[1] ODA models involving a single optimal threshold will henceforth be called "fixed-threshold" models. This note proposes a new "relative-threshold" ODA model for an inter-examiner reliability study[2] in which four examiners independently rate teeth condition for a sample of ten patients: "An important inferential question is whether the rater effects differ significantly from one another" (p. 19). In the original study, analysis of variance showed rater C assigned the greatest mean rating across patients: "The inference is therefore drawn that differential measurement bias exists (i.e., the *k* examiners differ systematically from one another in their mean levels of measurement)" (pp. 20-21). ODA was used to compare the entire response distribution (not only means) between raters. A fixed-threshold model identified no effects. A relative-threshold model tested the hypothesis that, for each observation in the sample considered separately, the rating by rater X will be less than (or equal to) the rating by rater Y. Analysis showed that the distribution of ratings made by rater C was nearly perfectly greater than corresponding (non-discriminable) ratings made by raters A, B, and D. This finding hints of possible development of optimal analogues of multidimensional scaling[3] and facet theory[4] methodologies.

Ratings (Table 1) were compared across patients between raters by ODA using fixed discriminant threshold values.[5-7] Analysis comparing all four raters simultaneously yielded a training model with relatively weak effect strength (ESS=16.7), that was theoretically insubstantial (D=20.0) and statistically insignificant (p<0.99). This result indicates that numerical ratings of the ten patients couldn't be discriminated between the four raters—thus no measurement bias exists.

2155-0182/10/$3.00

Table 1: Reliability Study Data: DMFS Scores of Four Raters for Ten Patients[2]

	Rater			
Patient	A	B	C	D
1	8	7	11	7
2	13	11	15	13
3	0	0	2	1
4	3	6	9	6
5	13	13	17	10
6	19	23	27	18
7	0	0	1	0
8	2	0	4	5
9	18	20	22	16
10	5	3	8	3

Results of analyses of ratings made by six unique parings of four independent raters were consistent with omnibus results: training ODA models had moderate to relatively weak effect strength ($10 \leq ESS \leq 30$), were theoretically marginal (D's> 4.7), and statistically insignificant (p's>0.739).

Evaluated using a fixed discriminant threshold applied to the sample, ratings made for these ten patients couldn't be discriminated between any of the four raters. That is, analysis which maximized the classification accuracy in discriminating raters on the basis of their ratings of patients indicates that the four raters assigned statistically comparable distributions of ratings across the ten patients. This finding suggests that no measurement bias exists.

Relative Discriminant Threshold

For each of six different rater pairings, patient ratings were compared between rater by ODA using relative discriminant threshold values. For clarity of exposition, Table 2 demonstrates this new methodology for comparing pairs of raters.

Consider first the comparison of raters A and B. Two relative threshold criteria are: (1) A is less than or equal to B; and (2) A is less than B (any observation receiving the identical rating from raters A and B is deleted from the sample). To change the direction of an inequality (i.e., $C \leq D$ *vs.* $D \leq C$), simply change "0" to "1" (and vice versa) in Table 2.

As seen in Table 2, by the first criterion 6 of 10 patients received a rating from rater A that was as low/lower than the corresponding rating from rater B. If it is assumed rating is a uniform random variable then p(success)=0.50 on a given trial, and the binomial probability of 6 successes in 10 trials is $p<0.206$. And, by the second criterion, 3 of 6 patients had a rater A rating that was lower than the corresponding rater B rating ($p<0.274$). Comparable results emerged for the comparisons of raters A and D (p's<0.247 and 0.219 for the first and second criteria), and of raters B and D (p's<0.118 and 0.313 for the first and second criteria).

Table 2: Comparing Paired Raters by Relative Threshold: 0 = No; 1 = Yes; --- = Missing

	Rater		Relative Threshold	
Patient	A	B	$A \leq B$	$A < B$
1	8	7	0	0
2	13	11	0	0
3	0	0	1	---
4	3	6	1	1
5	13	13	1	---
6	19	23	1	1
7	0	0	1	---
8	2	0	0	0
9	18	20	1	1
10	5	3	0	0
Patient	A	C	$A \leq C$	$A < C$
1	8	11	1	1
2	13	15	1	1
3	0	2	1	1
4	3	9	1	1
5	13	17	1	1
6	19	27	1	1
7	0	1	1	1
8	2	4	1	1
9	18	22	1	1
10	5	8	1	1

2155-0182/10/$3.00

Patient	A	D	A≤D	A<D
1	8	7	0	0
2	13	13	1	---
3	0	1	1	1
4	3	6	1	1
5	13	10	0	0
6	19	18	0	0
7	0	0	1	---
8	2	5	1	1
9	18	16	0	0
10	5	3	0	0
Patient	B	C	B≤C	B<C
1	7	11	1	1
2	11	15	1	1
3	0	2	1	1
4	6	9	1	1
5	13	17	1	1
6	23	27	1	1
7	0	1	1	1
8	0	4	1	1
9	20	22	1	1
10	3	8	1	1
Patient	B	D	B≤D	B<D
1	7	7	1	---
2	11	13	1	1
3	0	1	1	1
4	6	6	1	---
5	13	10	0	0
6	23	18	0	0
7	0	0	1	---
8	0	5	1	1
9	20	16	0	0
10	3	3	1	---
Patient	C	D	C≤D	C<D
1	11	7	0	0
2	15	13	0	0
3	2	1	0	0
4	9	6	0	0
5	17	10	0	0
6	27	18	0	0
7	1	0	0	0
8	4	5	1	1
9	22	16	0	0
10	8	3	0	0

In contrast, by both the first and second criteria, all 10 patients received a rating from rater C which was higher than the corresponding rating from raters A or B (p's<0.0010), and 9 of 10 patients (all but patient #8) received a rating from rater C higher than the corresponding rating from rater D (p<0.0098). While patient ratings made by raters A, B and D could not be discriminated from each other, corresponding ratings made by rater C were significantly—and almost unilaterally—higher than made by all the other raters. The uniform (constant) character of the difference between ratings made by rater C *vs*. other raters suggests ipsative standardization of each rater's ratings should be used in order to circumvent possible confounding attributable to mean differences between raters.[6,8] It should be noted that the comparability of distributions of raters' ratings over a sample doesn't address if or how raters' ratings are related to each other, but consideration of inter-rater reliability[9-12] lies outside the purview of this note.

Use of an adaptive, sequentially applied ordinal inequality—instead of a fixed value—as an optimal threshold is consistent with other ODA methods developed for temporal (time-ordered) designs, such as little-jiffy[13,14], single-case series[15], and weighted Markov[16,17] applications. However this note is the first to report that such adaptive mechanisms may also be used in single-point-in-time applications such as is the case with paired comparisons.

References

[1]Yarnold PR (2017). What is optimal data analysis? *Optimal Data Analysis*, *6*, 26-42.

[2]Fleiss JL (1986). *The design and analysis of clinical experiments*. New York, NY: Wiley (pp. 19-26).

[3]Schiffman SS, Reynolds ML, Young FW (1981). *Introduction to multidimensional scaling: Theory, methods, and applications.* Orlando, FL: Academic Press.

2155-0182/10/$3.00

[4]Brog I, Shye S (1995). *Facet theory: Form and content*. Thousand Oaks, CA: Sage.

[5]Yarnold PR, Soltysik RC (1991). Theoretical distributions of optima for univariate discrimination of random data. *Decision Sciences*, *22*, 739-752.

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

[7]Yarnold PR (2018). Visualizing application and summarizing accuracy of ODA models. *Optimal Data Analysis, 7*, 85-89.

[8]Yarnold PR (1996). Characterizing and circumventing Simpson's paradox for ordered bivariate data. *Educational and Psychological Measurement, 56*, 430-442.

[9]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.

[10]Yarnold PR (2016). Matrix display of pairwise novometric associations for ordered variables. *Optimal Data Analysis*, *5*, 94-101.

[11]Yarnold PR, Linden A (2016). Novometric analysis with ordered class variables: The optimal alternative to linear regression analysis, *Optimal Data Analysis*, *5*, 65-73.

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

[13]Yarnold PR (2013). Determining when annual crude mortality rate most recently began increasing in North Dakota counties, I: Backward-stepping little jiffy. *Optimal Data Analysis*, 2, 217-219.

[14]Yarnold PR (2013). The most recent, earliest, and *K*th significant changes in an ordered series: Traveling backwards in time to assess when annual crude mortality rate most recently began increasing in McLean County, North Dakota. *Optimal Data Analysis*, *2*, 143-147.

[15]Yarnold PR (2013). Surfing the *Index of Consumer Sentiment*: Identifying statistically significant monthly and yearly changes. *Optimal Data Analysis*, *2*, 211-216.

[16]Yarnold PR, Soltysik RC (2019). Confirming the efficacy of weighting in optimal Markov analysis: Modeling serial symptom ratings. *Optimal Data Analysis*, *8*, 53-55.

[17]Yarnold PR (2019). Optimal Markov model relating *two* time-lagged outcomes. *Optimal Data Analysis, 8*, 61-63.

Author Notes

No conflict of interest was reported.