Optimal Data Analysis
Vol. 10 (March 16, 2021), 12-17

2155-0182/10/$3.00

Implementing ODA from Within Stata: Confirmatory and Exploratory Inter-Rater Reliability Hypothesis with a Three-Category Ordinal Rating

Paul R. Yarnold, Ph.D. and Ariel Linden, Dr.P.H.
Optimal Data Analysis, LLC and Linden Consulting Group, LLC

This paper illustrates testing directional (confirmatory) and non-directional (exploratory) hypotheses for an inter-rater reliability study using a three-category ordinal measure, via the Stata package for implementing ODA.

Recent papers[1-23] introduce the new Stata package called **oda**[24] for implementing ODA from within the Stata environment. This package is a wrapper for the MegaODA software system[25-27], so the MegaODA.exe file must be loaded on the computer for the **oda** package to work.[28] To download the **oda** package, at the Stata command line type: “ssc install oda” (without the quotation marks). This paper demonstrates use of the **oda** package to evaluate directional and non-directional hypotheses for a design in which two independent cardiologists each use a three-level ordinal rating scale to assess the same set of 200 electrocardiograms.

Methods

Data

Woolson[29] presented hypothetical data on two cardiologists who independently classified a set of 200 electrocardiograms into one of three mutually exclusive and exhaustive ordinal diagnoses: normal (1), possibly abnormal (2), and abnormal (3).

Analytic Process

We first test the directional (“confirmatory”) alternative hypothesis that ratings made by the cardiologists are consistent—that is, fall into the major diagonal of the cross-classification table. Thus, the ratings made by cardiologist “X” (for lack of a better name) are directly discriminable (predictable) on the basis of ratings made by cardiologist “Y,” and *vice versa*. The null hypothesis is that this is not true.[30-36] Analysis was accomplished using the following **oda** syntax (see the help file for **oda** for a complete description of syntax options):

2155-0182/10/$3.00

```
oda ratingx ratingy,
pathoda("C:\ ODA\")
store("C:\Users\Ariel\Desktop\ODA\")
iter(25000) direction(< 1 2 3)
```

This syntax is explained as follows. Here "ratingx" is the *class* variable and "ratingy" is the *attribute*. The order of these two variables could be reversed since they both have the same number of levels (three), and there is no specific difference between the raters. But, if one rater was an expert, and the other rater was a trainee, a researcher might be interested in ascertaining how well the novice (treated as the attribute) was able to correctly apply the same ratings as the expert (treated as the class variable).

Here, "C:\ODA\" is the directory path where the MegaODA.exe file exists on the computer, and where other files generated in analysis are stored; 25,000 iterations (repetitions) are used to obtain a permutation *p*-value; and the directional hypothesis is that the raters' ratings agree.[23,24] The **oda** package produces an extract of the total output produced by ODA software (the complete output is stored in the specified directory with the extension ".out").

```
ODA model:
----------
IF RATINGY <= 1.5 THEN RATINGX = 1
IF 1.5 < RATINGY <= 2.5 THEN RATINGX = 2
IF 2.5 < RATINGY THEN RATINGX = 3

Summary for Class RATINGX  Attribute RATINGY
--------------------------------------------

Performance Index          Train
-----------------         ------
Overall Accuracy           65.00%
PAC RATINGX=1              75.00%
PAC RATINGX=2              50.00%
PAC RATINGX=3              50.00%
Effect Strength PAC        37.50%
PV RATINGX=1               90.00%
PV RATINGX=2               33.33%
PV RATINGX=3               50.00%
Effect Strength PV         36.67%
Effect Strength Total      37.08%

Monte Carlo summary (Fisher randomization):
-------------------------------------------
Iterations:  25000
Estimated p: 0.000000
```

Effect strength for sensitivity (ESS) is labelled in the output as "Effect Strength PAC" (Percentage Accurate Classification). For the exploratory hypothesis ESS is 37.5%, which exceeds the minimum criterion (ESS≥25) to be classified as a moderate effect.[30] This result is statistically significant: $p<0.0001$ (this is conventional reporting: to be precise, as there were 25,000 iterations, $p<1/25000$, or $p<0.00004$).

The directional analysis just conducted tests the a priori hypothesis that rater's ratings exactly agree. However, one may also evaluate the exploratory hypothesis that rater's ratings agree, but in a discordant manner. For example, a rating of 1 for rater X corresponds to a rating of 2 for rater Y, and *vice versa*.[32] This hypothesis is evaluated using a non-directional analysis, and the oda code is identical to the code given earlier, except that the directional command [i.e., direction(< 1 2 3)] is deleted. When this analysis was conducted the results were the same as obtained for the confirmatory analysis.

We believe ODA should be considered the preferred statistical approach *vs*. alternative methods since it avoids statistical assumptions required of conventional models, is insensitive to skewed data or outliers, and has the ability to handle any variable metric including categorical, Likert-type integer, and real number measurement scales.[30] In contrast to alternative methods, only ODA can identify the optimal (maximum-accuracy) assignments (categorical attributes) or cutpoints (ordered attributes) that exist for the attribute, which in turn facilitates the use of measures of predictive accuracy.

Furthermore, ODA can evaluate model reproducibility by multiple methods, allowing assessment of potential cross-generalizability of the model when it is applied to classify independent random samples.[30]

For these reasons we recommend that researchers employ ODA and CTA frameworks to evaluate the statistical hypotheses which are explored in their laboratory and field research endeavors.[37-56]

2155-0182/10/$3.00

References

[1]Linden A (2020). Implementing ODA from within Stata: An application to data from a randomized controlled trial (*Invited*). *Optimal Data Analysis, 9*, 9-13.

[2]Linden A (2020). Implementing ODA from within Stata: Implementing ODA from within Stata: An application to estimating treatment effects using observational data (*Invited*). *Optimal Data Analysis, 9*, 14-20.

[3]Linden A (2020). Implementing ODA from within Stata: An application to dose-response relationships (*Invited*). *Optimal Data Analysis, 9*, 26-32.

[4]Linden A (2020). Implementing ODA from within Stata: assessing covariate balance in observational studies (*Invited*). *Optimal Data Analysis, 9*, 33-38.

[5]Linden A (2020). Implementing ODA from within Stata: Evaluating treatment effects for survival (time-to-event) outcomes (*Invited*). *Optimal Data Analysis*, *9*, 39-44.

[6]Linden A (2020). Implementing ODA from within Stata: Evaluating treatment effects in multiple-group interrupted time series analysis (Invited). *Optimal Data Analysis*, *9*, 45-50.

[7]Linden A (2020). Implementing ODA from within Stata: identifying structural breaks in single-group interrupted time series designs (Invited). *Optimal Data Analysis*, *9*, 51-56.

[8]Linden A (2020). Implementing ODA from within Stata: Finding the optimal cut-point of a diagnostic test or index (Invited). *Optimal Data Analysis*, *9*, 74-78.

[9]Yarnold PR, Linden A (2020). Implementing ODA from within Stata: Exploratory hypothesis, binary class variable, and binary attribute. *Optimal Data Analysis*, *9*, 94-98.

[10]Yarnold PR, Linden A (2020). Implementing ODA from within Stata: Confirmatory hypothesis, binary class variable, and binary attribute. *Optimal Data Analysis*, *9*, 99-103.

[11]Yarnold PR, Linden A (2020). Implementing ODA from within Stata: Exploratory hypothesis, binary class variable, and binary attribute. *Optimal Data Analysis*, *9*, 104-108.

[12]Yarnold PR, Linden A (2020). Implementing ODA from within Stata: Exploratory hypothesis, binary class variable, and ordinal (rank) attribute. *Optimal Data Analysis*, *9*, 109-113.

[13]Yarnold PR, Linden A (2020). Implementing ODA from within Stata: confirmatory hypothesis, binary class variable, and ordinal attribute. *Optimal Data Analysis*, *9*, 128-132.

[14]Yarnold PR, Linden A (2020). Implementing ODA from within Stata: Exploratory hypothesis, binary class variable, categorical ordinal attribute. *Optimal Data Analysis*, *9*, 133-136.

[15]Yarnold PR, Linden A (2020). Implementing ODA from within Stata: Nondirectional hypothesis, binary class variable, categorical ordinal attribute. *Optimal Data Analysis*, *9*, 137-140.

[16]Yarnold PR, Linden A (2020). Implementing ODA from within Stata: Directional hypothesis, binary class variable, ordinal attribute. *Optimal Data Analysis*, *9*, 141-145.

[17]Yarnold PR, Linden A (2020). Implementing ODA from within Stata: Confirmatory hypothesis, binary class variable, continuous attribute. *Optimal Data Analysis*, *9*, 146-151.

[18]Yarnold PR, Linden A (2020). Implementing ODA from within Stata: Nondirectional, multicategorical class variable, multicategorical attribute. *Optimal Data Analysis*, *9*, 152-156.

2155-0182/10/$3.00

[19]Yarnold PR, Linden A (2020). Implementing ODA from within Stata: Directional hypothesis, multicategorical class variable and attribute. *Optimal Data Analysis*, *9*, 157-161.

[20]Yarnold PR, Linden A (2020). Implementing ODA from within Stata: Directional hypothesis, multicategorical class variable, ordinal attribute. *Optimal Data Analysis*, *9*, 162-166.

[21]Yarnold PR, Linden A (2020). Implementing ODA from within Stata: *A Priori* hypothesis, three-category class variable, four-level (integer) attribute. *Optimal Data Analysis*, *9*, 167-171.

[22]Linden A, Yarnold PR (2020). Implementing ODA from within Stata: A reanalysis of the National Supported Work Experiment. *Optimal Data Analysis*, *9*, 178-182.

[23]Yarnold PR, Linden A (2021). Implementing ODA from within Stata: Exploratory hypothesis, three-category class variable, continuous attribute. *Optimal Data Analysis*, *10*, 3-9.

[24]Linden A (2020). ODA: Stata module for conducting Optimal Discriminant Analysis. *Statistical Software Components S458728, Boston College Department of Economics*.

[25]Soltysik RC, Yarnold PR (2013). MegaODA large sample and BIG DATA time trials: Separating the chaff. *Optimal Data Analysis*, *2*, 194-197.

[26]Soltysik RC, Yarnold PR (2013). MegaODA large sample and BIG DATA time trials: Harvesting the Wheat. *Optimal Data Analysis*, *2*, 202-205.

[27]Yarnold PR, Soltysik RC (2013). MegaODA large sample and BIG DATA time trials: Maximum velocity analysis. *Optimal Data Analysis*, *2*, 220-221.

[28]Rhodes NJ, Yarnold PR. 2020. ODA: a package and R-interface for the MegaODA software suite. R package version 1.0.1.3. Available: https://github.com/njrhodes/ODA

[29]Woolson RF (1987). *Statistical methods for the analysis of biomedical data*. New York, NY: Wiley.

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

[31]Bryant FB, Harrison PR (2013). How to create an ASCII input data file for UniODA and CTA software (*Invited*). *Optimal Data Analysis, 2*, 2-6.

[32]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.

[33]Yarnold PR (2015). Estimating inter-rater reliability using pooled data induces paradoxical confounding: An example involving Emergency Severity Index triage ratings. *Optimal Data Analysis*, *4*, 21-23.

[34]Yarnold PR (2015). UniODA *vs*. Spearman rank ρ: Between-raters reliability of scores on the Adverse Drug Reaction Probability Scale. *Optimal Data Analysis*, *4*, 148-150.

[35]Yarnold PR (2016). Novometric *vs*. ODA reliability analysis *vs*. polychoric correlation with relaxed distributional assumptions: Inter-rater reliability of independent ratings of plant health. *Optimal Data Analysis*, *5*, 179-183.

[36]Yarnold PR (2019). Regression *vs*. novometric-based assessment of inter-examiner reliability. *Optimal Data Analysis*, *8*, 107-111.

2155-0182/10/$3.00

[37]Linden A, Yarnold PR, Nallomothu BK (2016). Using machine learning to model dose-response relationships. *Journal of Evaluation in Clinical Practice*, *22*, 860-867.

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

[39]Yarnold PR, Linden A (2016). Theoretical aspects of the D statistic. *Optimal Data Analysis*, *22*, 171-174.

[40]Linden A, Yarnold PR (2017). Using classification tree analysis to generate propensity score weights. *Journal of Evaluation in Clinical Practice*, *23,* 703-712.

[41]Linden A, Yarnold PR (2017). Modeling time-to-event (survival) data using classification tree analysis. *Journal of Evaluation in Clinical Practice*, *23*, 1299-1308.

[42]Linden A, Yarnold PR (2018). Identifying causal mechanisms in health care interventions using classification tree analysis. *Journal of Evaluation in Clinical Practice*, *24*, 353-361.

[43]Linden A, Yarnold PR (2017). Minimizing imbalances on patient characteristics between treatment groups in randomized trials using classification tree analysis. *Journal of Evaluation in Clinical Practice*, *23*, 1309-1315.

[44]Linden A, Yarnold PR (2018). Estimating causal effects for survival (time-to-event) outcomes by combining classification tree analysis and propensity score weighting. *Journal of Evaluation in Clinical Practice*, *24*, 380-387.

[45]Linden A, Yarnold PR (2016). Using machine learning to assess covariate balance in matching studies. *Journal of Evaluation in Clinical Practice*, 22, 848-854.

[46]Linden A, Yarnold PR (2016). Using machine learning to identify structural breaks in single-group interrupted time series designs. *Journal of Evaluation in Clinical Practice*, *22*, 855-859.

[47]Linden A, Yarnold PR (2016). Combining machine learning and matching techniques to improve causal inference in program evaluation. *Journal of Evaluation in Clinical Practice*, *22*, 868-874.

[48]Linden A, Yarnold PR (2016). Combining machine learning and propensity score weighting to estimate causal effects in multivalued treatments. *Journal of Evaluation in Clinical Practice*, *22,* 875-885.

[49]Linden A, Yarnold PR (2018). Using machine learning to evaluate treatment effects in multiple-group interrupted time series analysis. *Journal of Evaluation in Clinical Practice*, *24*, 740-744.

[50]Rhodes NJ (2020). Statistical power analysis in ODA, CTA and Novometrics (Invited). *Optimal Data Analysis*, *9*, 21-25.

[51]Yarnold PR, Brofft GC (2013). ODA range test *vs*. one-way analysis of variance: Comparing strength of alternative line connections. *Optimal Data Analysis, 2*, 198-201.

[52]Yarnold PR (2013). ODA range test *vs*. one-way analysis of variance: Patient race and lab results. *Optimal Data Analysis, 2*, 206-210.

[53]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

2155-0182/10/$3.00

[54]Yarnold PR (2016) Causality of adverse drug reactions: The upper-bound of arbitrated expert agreement for ratings obtained by WHO and Naranjo algorithms. *Optimal Data Analysis*, *5*, 37-40.

[55]Yarnold PR (2016). Novometric *vs*. ODA reliability analysis *vs*. polychoric correlation with relaxed distributional assumptions: Inter-rater reliability of independent ratings of plant health. *Optimal Data Analysis, 5*, 179-183.

[56]Rhodes NJ, Yarnold PR (2020). Generating novometric confidence intervals in R: Bootstrap analyses to compare model and chance ESS. *Optimal Data Analysis, 9*, 172-177.

Author Notes

No conflicts of interest were reported.