Optimal Data Analysis
Vol. 7 (March 4, 2018), 16-18

2155-0182/10/$3.00

Friedman Test *vs*. ODA *vs*. Novometry: Rating Violin Excellence

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

The Friedman test, ODA, and novometric statistical analysis conducted via ODA, are used to compare ten expert violinists' independent, blind evaluations of three different violins, each rated using 10-point scales.

A non-parametric alternative to one-way ANOVA with repeated measures, the Friedman test is used to test for between-group differences when the dependent variable is ordinal, or when the data are continuous but violate assumptions required by one-way ANOVA with repeated measures.[1]

Data used for exposition are ten expert violinists' independent, blind ratings of the "excellence" of three different violins. Ratings were made on a ten-point scale: 1="lowest rating", 10= "highest rating" (Table 1).

Ratings of the three violins were transformed into ranks separately by violinist: a rank of "3" is assigned to the largest of the violinist's three ratings, a rank of "2" to the intermediate rating, and a rank of "1" to the smallest of the three ratings made by the violinist.

For the Friedman test, chi-square (df=2, N=30)= 9.95, $p<0.01$, indicating: "…observed differences among the mean rankings for the three violins reflect something more than mere random variability, something more than mere chance coincidence among the judgments of the expert players".[2]

Analyses required to disentangle the omnibus finding were not reported.

Table 1: Independent, Blind Ratings (the Attribute) of Three Violins (the Class Variable) Made By Ten Expert Violinists

Violinist	Violin A	Violin B	Violin C
1	9.0	7.0	6.0
2	9.5	6.5	8.0
3	5.0	7.0	4.0
4	7.5	7.5	6.0
5	9.5	5.0	7.0
6	7.5	8.0	6.5
7	8.0	6.0	6.0
8	7.0	6.5	4.0
9	8.5	7.0	6.5
10	6.0	7.0	3.0

Omnibus ODA Analysis

Herein, 25,000 Monte Carlo experiments were used to estimate *p*; leave-one-out (LOO) single-sample jackknife analysis was performed; and ESS was evaluated using conventional rule-of-thumb criteria: ESS<25 is a relatively weak effect, ESS<50 is a moderate effect, and ESS <75 is a relatively strong effect.[3]

For ODA, ratings were treated as an ordered attribute, and were compared between

2155-0182/10/$3.00

three violins—treated as a three-category class variable.[3] The ODA model that emerged was:

If Rating ≤ 6.25 predict Violin C;

If 6.25 < Rating < 7.25 predict Violin B;

If 7.25 < Rating predict Violin A.

This model indicates Violin C was rated as being least excellent; Violin A as being most excellent; and Violin B as having an intermediate level of excellence. In training (total sample) analysis this model yielded high-moderate predictive accuracy (ESS=45.0, $p< 0.025$), with inter-rater consensus of 70% for Violin A, and 60% for Violins B and C. Predictive accuracy fell to a moderate level of ESS= 35% in LOO analysis because inter-rater consensus fell to 40% for Violin B.

Pairwise ODA Analyses

To further explore differences between violins, excellence ratings were compared between the three different *pairings* of violins.[3]

When comparing Violins A and B the ODA model that emerged was:

If Rating ≤ 7.25 predict Violin B;

If 7.25 < Rating predict Violin A.

This model indicates that Violin B was rated as being less excellent than Violin A. The effect was relatively strong (ESS=50.0) but was statistically marginal ($p<0.065$) due to the small sample size. Nevertheless, there was inter-rater consensus of 70% for Violin A, and 80% for Violin B. The effect was stable in LOO analysis, with one-tailed $p<0.035$ (statistically significant if evaluated by the generalized per-comparison criterion, but not if evaluated by the experimentwise criterion).

When comparing Violins B and C the ODA model that emerged was:

If Rating ≤ 6.25 predict Violin C;

If 6.25 < Rating predict Violin B.

This model indicates that Violin C was rated as being less excellent than Violin B. The effect was moderate (ESS=40.0) but was not statistically reliable ($p<0.31$) due to the small sample size. There was inter-rater consensus of 80% for Violin B, and 60% for Violin C. ESS fell to 20.0 in LOO analysis.

Finally, comparing Violins A and C the ODA model that emerged was:

If Rating ≤ 6.75 predict Violin C;

If 6.75 < Rating predict Violin A.

This model indicates that Violin C was rated as being less excellent than Violin A. The effect was relatively strong (ESS=60.0) and statistically reliable ($p<0.049$). There was inter-rater consensus of 80% for Violins A and C in training analysis, but consensus fell to 70% for Violin C in LOO analysis (ESS=50.0; $p<0.035$).

Considered as a whole, the pairwise comparisons showed that only the difference in excellence ratings between Violins A and C was statistically reliable and relatively strong.

Novometric Analysis

Novometric analysis requires assessing the ESS yielded by all possible rules (models) predicting class categories (Violins A, B, and C).[4,5] The first required comparison was the omnibus ODA analysis conducted earlier: Violin A *vs*. Violin B *vs*. Violin C.

The remaining comparisons to be evaluated are: Violins A and B *vs*. Violin C; Violins A and C *vs*. Violin B; and Violin A *vs*. Violins B and C.

For the comparison (A and B) *vs*. C, the ODA model that emerged was:

If Rating ≤ 6.75 predict Violin C;

If 6.75 < Rating predict Violins A and B.

2155-0182/10/$3.00

This model yielded relatively strong inter-rater consensus (ESS=50.0) which was statistically significant (p<0.036) and stable in LOO analysis (p<0.014). There was inter-rater consensus of 70% for Violins A and B, and 80% for Violin C.

For the comparison (A and C) *vs*. B, the ODA model was relatively weak (ESS=20.0), statistically unreliable (p<0.86), and unstable in LOO analysis (ESS=0).

Finally, for the comparison A *vs*. (B and C), the ODA model that emerged was:

If Rating ≤ 7.25 predict Violins B and C;

If 7.25 < Rating predict Violin A.

This model yielded relatively strong inter-rater consensus (ESS=55.0) that was statistically significant (p<0.015) and stable in LOO analysis (p<0.0049). There was inter-rater consensus of 70% for Violin A, and 85% for Violins B and C.

Table 2 summarizes the ESS obtained by the competing novometric models.

Table 2: Training and LOO ESS of Novometric Models

Comparison	Training	LOO
A *vs*. B *vs*. C	45	35
(A, B) *vs*. C	50	50
(A, C) *vs*. B	20	0
A *vs*. (B, C)	55	55

ESS obtained in training analysis is an upper-bound estimate of expected predictive accuracy, due to capitalization on chance. ESS obtained in reproducibility analysis (e.g., LOO) is a more tempered estimate of effect strength. In the present example the novometric (globally optimal) model is A *vs*. (B, C), which has the strongest ESS in LOO of all possible models (Table 2). This model indicates that Violin A is more excellent than Violins B and C, which cannot be discriminated from one another.

As seen, novometric analysis of three-category multicategorical problems can sometimes be easily accomplished manually using UniODA or MegaODA software for applications involving samples which are too small to support CTA.

References

[1]Friedman M (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association, 32*, 675–701.

[2]http://vassarstats.net/textbook/ch15a.html

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

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

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

Author Notes

This article reanalyzes publically-available data and is exempt from IRB review. No conflict of interest was reported.