Optimal Data Analysis
Vol. 5 (May 14, 2016), 19-23

2155-0182/10/$3.00

Pairwise Comparisons using UniODA *vs.* *Not* Log-Linear Model: Ethnic Group and Schooling in the 1980 Census

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

Data are from a contingency table used to determine the relationship between years of schooling arbitrarily parsed into six ordered categories, and ethnic group measured on a categorical variable with seven levels.[1] Although ordinal data are inappropriate for analysis via chi-square-based methods, log-linear analysis was used to investigate association between years of schooling and ethnic group.[2-10] Because the independence model didn't provide an acceptable representation of the data, it is clear that some form of association underlies the data. A three-dimensional log-linear-based solution was proposed: "In terms of the scores in the first dimension only, whites are closest to Chinese; blacks are closest to Vietnamese; and Hispanics are extreme outliers. Either the distance matrix…or a two-dimensional plot…can be used to locate the groups or measure distances in terms of educational distributions" (p. 103).[1] All possible pairwise comparisons were conducted between ethnic groups using UniODA, and the results revealed a single dimension (years of schooling) perfectly described the statistical conclusions reached for 20 of 21 analyses.[8,9] The single inconsistent finding had an associated miniscule effect size.

The data (frequencies) are given in Table 1: two ethnic groups included in the original data but omitted here are blacks (a conglomeration of foreign and domestic people) and Hispanic (which is independent of ethnicity).[1] All possible optimal (maximum-accuracy) pairwise comparisons of years of schooling between ethnic groups were obtained by conducting 21 separate analyses via the following MegaODA[11-13] command syntax (the sample was too large for UniODA software[8]):

Table 1: Years of Schooling and Ethnic Group[1]

	Years of Schooling					
Ethnicity	0	<12	12	<16	16	>16
White	57	4752	6310	3919	1545	1248
Japanese	105	3997	8312	7676	3775	2965
Chinese	1088	6359	4431	7007	3995	5546
Filipino	153	5054	3647	5408	4038	3478
Korean	206	2539	2925	2494	1980	1237
Indian	161	1981	1497	2423	1637	4699
Vietnamese	260	2785	1481	2231	344	352

2155-0182/10/$3.00

```
OUTPUT example.out;
OPEN example.dat;
VARS ethnic years;
CLASS ethnic;
ATTR years;
MC ITER 25000;
LOO;
GO;
```

The findings of the 21 UniODA pairwise comparisons of years of schooling conducted between pairs of ethnic groups are summarized in Table 2.

Table 2: Findings of 21 Pairwise Comparisons of Years of Schooling for Pairs of Ethnic Groups

	White	Japanese	Chinese	Filipino	Korean	Indian
Japanese	≤ 12 → W 62.4%, 53.7% ESS = 16.1 W < J					
Chinese	≤ 12 → W 62.4%, 58.2% ESS = 20.6 W < C	< 12 → C 84.7%, 26.2% ESS = 10.9 C < J				
Filipino	≤ 12 → W 62.4%, 59.3% ESS = 21.7 W < F	< 16 → J 74.9%, 34.5% ESS = 9.4 J < F	≤ 16 → F 19.5%, 84.0% ESS = 3.5 F < C			
Korean	< 16 → W 84.3%, 28.3% ESS = 12.6 W < K	< 12 → K 84.7%, 24.1% ESS = 8.8 K < J	≤ 16 → K 19.5%, 89.1% ESS = 8.6 K < C	< 12 → K 59.3%, 49.8% ESS = 9.2 K < F		
Indian	< 16 → W 84.3%, 51.1% ESS = 35.4 W < I	≤ 16 → J 89.0%, 37.9% ESS = 26.8 J < I	≤ 16 → C 80.5%, 37.9% ESS = 18.4 C < I	≤ 16 → F 84.0%, 37.9% ESS = 21.9 F < I	≤ 16 → K 89.1%, 37.9% ESS = 27.0 K < I	
Vietnamese	< 12 → V 73.0%, 40.9% ESS = 13.9 V < W	< 12 → V 84.7%, 40.9% ESS = 25.6 V < J	< 16 → V 33.6%, 90.7% ESS = 24.2 V < C	< 16 → V 34.5%, 90.7% ESS = 25.2 V < F	< 16 → V 28.3%, 90.7% ESS = 18.9 V < K	< 16 → V 51.1%, 90.7% ESS = 41.8 V < I

Note: All pairwise comparisons met the Sidak experimentwise criterion for statistical significance ($p < 0.05$), and all models had stabile ESS (0 = the level of predictive accuracy expected by chance, 100 = perfect prediction) in training (total sample) and leave-one-out (jackknife validity) analysis.[8,9,14-17] Discussed in text, within each cell of the table the first row is the UniODA model; the second row gives the percent of the column, and row ethnic group correctly classified by the model; the third row gives the normed predictive accuracy (ESS) of the model; and the fourth row presents a symbolic representation of the statistical analysis conclusion for the indicated comparison.

2155-0182/10/$3.00

Within each cell of the table the first row gives the UniODA model. For example, for the entry in the cell corresponding to column = Chinese, row = Korean, the UniODA model is: if years of schooling ≤ 16 then predict the observation is Korean; otherwise predict Chinese. The second row in each cell gives the percent of the ethnic group corresponding to the column, and the percent corresponding to the row, that was correctly classified by the model. For the example cell the UniODA model correctly classified 19.5% of Chinese and 89.1% of Korean observations. This indicates that the primary difference in this comparison is that proportionately more Koreans have 16 or fewer years of schooling, compared to Chinese. The third row in each cell gives ESS (the normed predictive accuracy) for the UniODA model. Finally, the fourth row in each cell is a symbolic representation of the statistical analysis conclusion for the indicated comparison.

Examining Individual Effects

Most effects were relatively weak, with the exception being pairwise comparisons involving Indian (four of six were of moderate strength) and Vietnamese (three of six were of moderate strength) observations. In models involving Indians the other ethnic group has disproportionately greater numbers of people with 16 or fewer years of schooling. Half of the models involving Vietnamese are dominated by the Vietnamese having disproportionately greater numbers of people with fewer than 12 years of schooling, or by the other group having disproportionately greater numbers of people with 16 or more years of schooling.

Identifying Structure Underlying the Pairwise Comparison Table

Methodology used to disentangle systems of pairwise comparisons is described elsewhere in the context of identifying unidimensional and multidimensional structure in Markov state transition tables and geologic core samples.[8]

In Table 3 an "X" indicates an effect in Table 2 that must be accounted for using the least complex (smallest dimensionality) model that can be identified.

Table 3: Effects to Account For at the Start of the Procedure

	W	J	C	F	K	I
J	X					
C	X	X				
F	X	X	X			
K	X	X	X	X		
I	X	X	X	X	X	
V	X	X	X	X	X	X

There are many ways to begin, but a method that can anchor the low (left-hand) and high (right-hand) ends of the model involves examining the results looking for a group that is always less than, or always greater than, the other groups. As seen in Table 2, scores for the Vietnamese observations are lower than scores for all other groups, and scores for white observations are lower than for all other groups other than Vietnamese observations. This pattern of results allows the left-hand (low schooling) side of the maximum-accuracy model to be mapped as illustrated in Figure 1.

Figure 1: Symbolic Representation of Maximum-Accuracy Model for W and V

V----W----

Eliminating columns and rows for V and W observations from Table 3 simplifies the pairwise comparisons disentanglement problem considerably (Table 4).

2155-0182/10/$3.00

Table 4: Effects Remaining After V and W

	J	C	F	K
C	X			
F	X	X		
K	X	X	X	
I	X	X	X	X

As seen in Table 2, scores for the Indian observations are greater than scores for all other groups, allowing the right-hand (high schooling) side of the maximum-accuracy model to be illustrated as seen in Figure 2.

Figure 2: Symbolic Representation of Maximum-Accuracy Model for W, V, and I

V----W---- // ----I

Eliminating row I and column K from Table 4 simplifies the pairwise comparisons disentanglement problem further (Table 5).

Table 5: Effects Remaining After V, W, and I

	J	C	F
C	C < J		
F	J < F	F < C	
K	K < J	K < C	K < F

Effects involving J, C, and K are easily mapped onto the schooling dimension, as seen in Figure 3.

Figure 3: Symbolic Representation of Model Involving J-C, J-K and C-K Comparisons

----K----C----J----

Eliminating the associated effects from Table 5 simplifies the pairwise comparisons disentanglement problem even more (Table 6).

Table 6: Final Effects Remaining to Model

	J	C	F
F	J < F	F < C	
K			K < F

As seen, the J < F and K < F effects sit at the right-hand-side of the series in Figure 3, and only the F < C effect is inconsistent with the unidimensional model that represents the other 20 pairwise comparisons in Table 2. The final maximum-accuracy model is seen in Figure 4, in which the single discordant pairwise comparison is indicated using red text.

Figure 4: Symbolic Representation of Final Maximum-Accuracy Model

V----W----K----C----J----F----I

In contrast to findings for the first of the three log-linear models that were identified in prior research, in the unidimensional UniODA model whites are closest to the Vietnamese and Korean observations, rather than to the Chinese observations. And, compared to the much more complex three-dimensional log-linear solution, the maximum-accuracy pairwise comparisons methodology explained (20 / 21) 95.2% of the pairwise effects using a single dimension—that is, years of schooling.

References

[1]Clogg CC, Shihadeh ES. *Statistical models for ordinal variables*. Thousand Oaks, CA: Sage, 1994 (pp. 101-103).

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

[3]Grimm LG, Yarnold PR (Eds.). *Reading and Understanding More Multivariate Statistics*. Washington, DC: APA Books, 2000.

2155-0182/10/$3.00

[4]Yarnold, P.R. (2010). UniODA *vs*. chi-square: Ordinal data sometimes feign categorical. *Optimal Data Analysis, 1*, 62-65.

[5]Yarnold, PR (2015). UniODA *vs. not* chi-square: Vaccine administration and flu. *Optimal Data Analysis, 4*, 159-160.

[6]Yarnold, PR (2015). UniODA *vs. not* chi-square: Work shift and raw material production quality. *Optimal Data Analysis, 4*, 168-170.

[7]Yarnold, PR (2015). UniODA-based structural decomposition *vs*. log-linear model: Statics and dynamics of intergenerational class mobility. *Optimal Data Analysis, 4*, 179-181.

[8]Yarnold PR, Soltysik RC. *Optimal data analysis: Guidebook with software for Windows*. Washington, DC: APA Books, 2005.

[9]Yarnold PR, Soltysik RC. *Maximizing predictive accuracy*. Chicago, IL: ODA LLC, 2016.

[10]Yarnold PR (2016). UniODA *vs. not* log-linear model: The relationship of mental health status and socioeconomic status. *Optimal Data Analysis*, *5*, 15-18.

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

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

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

[14]Linden A, Yarnold PR (2016). Using data mining techniques to characterize participation in observational studies. *Journal of Evaluation in Clinical Practice*. DOI: 10.1111/jep.12515

[15]Linden A, Yarnold PR (2016). Using machine learning to assess covariate balance in matching studies. *Journal of Evaluation in Clinical Practice*. DOI: 10.1111/jep.12538

[16]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*. DOI: 10.1111/jep.12544

[17]Linden A, Yarnold PR, Nallamothu BK (2016). Using machine learning to model dose-response relationships. *Journal of Evaluation in Clinical Practice*. DOI: 10.1111/jep.125xx

Author Notes

The study analyzed de-individuated data and was exempt from Institutional Review Board review. No conflict of interest was reported.

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