Optimal Data Analysis
Vol. 4 (December 9, 2015), 156-158

2155-0182/10/$3.00

UniODA *vs*. Chi-Square: Deciphering *R* x *C* Contingency Tables

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

Whereas omnibus effects identified using χ^2 to assess association in *R* x *C* contingency tables are often difficult or impossible to disentangle when *R* and *C* are greater than two, identifying the structure underlying *R* x *C* tables using UniODA is straightforward.[1] An investigation of the association between furniture factory production shift and type of defect is presented as an example.

Maximum-accuracy analysis of unweighted *R* x *C* tables is illustrated with an example in which a furniture company operating three production shifts categorized a total of *N* = 309 furniture defects into one of four qualitative categories[2] (an obvious weighted analysis would weight each defect by its associated cost, and the analysis would identify the model that explicitly minimized cost[1,3]). The resulting 3 x 4 contingency table is presented in Table 1.

Table 1: Production Shift and Type of Furniture Defect

	Production Shift			
Defect	1	2	3	Sum
A	15	26	33	74
B	21	31	17	69
C	45	34	49	128
D	13	5	20	38
Sum	94	96	119	309

The association between production shift and type of defect was assessed using chi-square analysis.[2] Here, χ^2 (df = 6, *N* = 309) = 19.2, $p < 0.05$. Therefore the null hypothesis that there is no association is rejected, and it is concluded that an association exists between shift and type of defect.[1] This finding begs two questions: (1) what, exactly, is the structure of the association that exists between production shift and type of defect, and (2) what is the strength of this association? Disentangling this omnibus effect and assessing effect size using chi-square in this example is challenging: intrigued readers are encouraged to decipher the finding.[4-6]

The nondirectional UniODA model discriminating production shift as a function of type of furniture defect was identified using the following UniODA[1,3] and MegaODA[7-9] software syntax:

```
OPEN factory.dat;
OUTPUT results.out;
VARS shift defect;
CLASS shift;
ATTRIBUTE defect;
CATEGORICAL defect;
```

2155-0182/10/$3.00

```
  MCARLO ITER 25000;
  GO;
```

The UniODA model was: if defect = A or D, then predict that shift = 3; if defect = B, then predict shift = 2; and if defect = C, then predict shift = 1. The association identified by UniODA was statistically significant ($p < 0.02$), however it reflects a relatively weak effect (*ESS* = 12.4).[1,3] The predictive accuracy of the model is summarized in the confusion table in Table 2.

Table 2: Confusion Table for UniODA Model

Actual	Predicted Shift			
Shift	1	2	3	
1	45	21	28	47.9
2	34	31	31	32.3
3	49	17	53	44.5
Predictive Value	35.2	44.9	47.3	

Model sensitivities reveal that defects associated with shifts 1 and 3 are more representative of the empirical data than are findings for shift 2. The model accurately classified 45 + 31 + 53 = 129 defects, leaving 309 – 129 = 180 defects unexplained. Eliminating the correctly classified observations from the original data creates a residual sample: secondary structure underlying the residuals can be identified by structural decomposition analysis using the same UniODA syntax with data in the residual sample used as the input data file.[1,3]

Statistical modeling is useful for identifying reliable effects, but in this example the effect size for the statistical model is weak, while the practical (logistical) implications of the model are complex—involving four interventions (one for each type of defect) for three shifts—all in an effort to eliminate 129 defects.

In light of the findings in Table 1, and in the absence of cost weights, a more reasonable approach presently would be to target the most likely defect—C, which accounts for 41.4% of all defects, and is the most frequently observed defect in shift 1 (47.9%), shift 2 (35.4%), and shift 3 (41.2%). This is not only less complex (one intervention used across all shifts), but it would facilitate assessing the generalizability of the intervention (and any specific requirements) across shifts. After the intervention is completed the third row of the original data is eliminated, simplifying the problem to a 3 x 3 contingency table. Assessing the efficacy of the intervention would also enable investigators to evaluate the possible interaction of the intervention with respect to the distribution of defects of all four qualitative types.

References

[1] Yarnold PR, Soltysik RC (In Review). *Maximizing predictive accuracy*. Chicago, IL: ODA Books.

[2] Mendenhall W, Reinmuth JE (1974). *Statistics for management and economics, 2nd Ed*. North Scituate, MA: Duxbury (p. 476).

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

[4] Bishop YMM, Fienberg SE, Holland PW (1978). *Discrete multivariate analysis: Theory and practice*. Cambridge, MA: MIT Press.

[5] Yarnold PR (2013). Analyzing categorical attributes having many response categories. *Optimal Data Analysis*, *2*, 172-176. URL: http://optimalprediction.com/files/pdf/V2A26.pdf

[6] Yarnold PR (2013). Univariate and multivariate analysis of categorical attributes with many response categories. *Optimal Data Analysis*, 2, 177-190. URL: http://optimalprediction.com/files/pdf/V2A27.pdf

[7]Soltysik RC, Yarnold PR (2013). MegaODA large sample and BIG DATA time trials: Separating the chaff. *Optimal Data Analysis*, 2, 194-197. URL: http://optimalprediction.com/files/pdf/V2A29.pdf

[8]Soltysik RC, Yarnold PR (2013). MegaODA large sample and BIG DATA time trials: Harvesting the Wheat. *Optimal Data Analysis,* 2, 202-205. URL: http://optimalprediction.com/files/pdf/V2A31.pdf

[9]Yarnold PR, Soltysik RC (2013). MegaODA large sample and BIG DATA time trials: Maximum velocity analysis. *Optimal Data Analysis*, 2, 220-221. URL: http://optimalprediction.com/files/pdf/V2A35.pdf

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