Optimal Data Analysis
Vol. 4 (November 27, 2015), 131-134

2155-0182/10/$3.00

Gender and Psychology Concentration for Graduate Students

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

Gender and psychology concentration are cross-classified for $N = 100$ graduate students, resulting in a 2 x 5 contingency table.[1] Estimating the nature and magnitude of the association between these variables by legacy statistical methods is inappropriate owing to computational and interpretative difficulties. However, obtaining and interpreting the maximum-accuracy model for these data is straightforward.

In Table 1 the minimum expectation for chi-square[2] is violated for cells with $N \leq 10$, undermining the validity of statistical analysis conclusions based on chi-square analysis, or on methods based on chi-square such as logistic regression[3,4] and log-linear[5,6] analysis. In addition, in this example computational and interpretative issues[1] also invalidate the use of legacy effect strength measures including ϕ^2, V^2, and T^2.

Table 1: Gender and Psychology Concentration for Graduate Students

	Gender	
Concentration	Male	Female
Educational	49	11
Social	2	8
Clinical	3	7
Experimental	1	9
Cognitive	0	10

Note: For maximum-accuracy analysis the data were coded[7]: for gender, female = 0, male = 1; for area, Educational = 1, Social = 2, Clinical = 3, Experimental = 4, Cognitive = 5.

The UniODA model for predicting (i.e., discriminating) gender (*gender*) on the basis of psychology concentration (*area*) was obtained via the following UniODA[8] and MegaODA[9-11] software syntax:

```
OPEN gender.dat;
OUTPUT gender.out;
VARS gender area;
CLASS gender;
ATTRIBUTE area;
CAT area;
MCARLO ITER 25000;
GO;
```

The CTA model[6,12-18] predicting gender on the basis of concentration was obtained via the identical syntax given above, but replacing the MCARLO command by the following CTA software[19] syntax:

```
MC ITER 5000 CUTOFF .05 STOP 99.9;
PRUNE .05;
ENUMERATE;
```

2155-0182/10/$3.00

Presently the UniODA[8], hierarchically-optimal[21], enumerated-optimal[22], and globally-optimal[23] CTA models were identical, and only one model emerged for the sample.[6,20,24,25] The maximum-accuracy (optimal) model was: if area = educational then predict gender = male; otherwise predict gender = female. The result was statistically significant (exact $p < 0.001$) and it reflected a relatively strong[6,8] effect (*ESS* = 65.6; *ESP* = 66.7).[26] The confusion table[6,8] summarizing model classification performance is presented in Table 2.

Table 2: Confusion Table for Maximum-Accuracy Model for Gender

		Predicted Gender	
		Female	Male
Actual	Female	34	11
Gender	Male	6	49

The model had strong sensitivity for classifying the female [34 / (34 + 11) x 100% = 75.6%] and the male [49 / (6 + 49) x 100% = 89.1%] graduate students. And, when making point predictions the model yielded strong predictive value for predicting female [34 / (34 + 6) x 100% = 85.0%] and male [49 / (49 + 11) x 100% = 81.7%] graduate students.

References

[1]Berry KJ, Johnston JE, Mielke PW (2007). An alternative measure of effect size for Cochran's *Q* test for related proportions. *Perceptual and Motor Skills*, *104*, 1236-1242.

[2]Yarnold JK (1970). The minimum expectation in χ^2 goodness of fit tests and the accuracy of approximations for the null distribution. *Journal of the American Statistical Association*, *65*, 864-886. URL: http://www.jstor.org/stable/2284594

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

[4]Yarnold PR (2014). UniODA *vs*. logistic regression analysis: Serum cholesterol and coronary heart disease and mortality among middle aged diabetic men. *Optimal Data Analysis*, *3*, 17-18. URL: http://optimalprediction.com/files/pdf/V3A6.pdf

[5]Grimm LG, Yarnold PR (2000). *Reading and understanding more multivariate statistics*. Washington, DC: APA Books.

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

[7]Bryant FB, Harrison PR (2013). How to create an ASCII input data file for UniODA and CTA software. *Optimal Data Analysis*, *2*, 2-6. URL: http://optimalprediction.com/files/pdf/V2A1.pdf

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

[9]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

[10]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

[11]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

2155-0182/10/$3.00

[12]Yarnold PR (1996). Discriminating geriatric and non-geriatric patients using functional status information: An example of classification tree analysis via UniODA. *Educational and Psychological Measurement, 56*, 656-667. DOI: 10.1177/0013164496056004007

[13]Yarnold PR, Soltysik RC, Bennett CL (1997). Predicting in-hospital mortality of patients with AIDS-related *Pneumocystis carinii* pneumonia: An example of hierarchically optimal classification tree analysis. *Statistics in Medicine, 16*, 1451-1463. DOI: 10.1002/(SICI)1097-0258(19970715)16:13<1451::AID-SIM571>3.0.CO;2-F

[14]Kanter AS, Spencer DC, Steinberg MH, Soltysik RC, Yarnold PR, Graham NM (1999). Supplemental vitamin B and progression to AIDS and death in black South African patients infected with HIV. *Journal of Acquired Immune Deficiency Syndrome, 21*, 252-253. URL: http://journals.lww.com/jaids/Citation/1999/07010/Supplemental_Vitamin_B_and_Progression_to_AIDS_and.11.aspx

[15]Kyriacou DN, Yarnold PR, Stein AC, Schmitt BP, Soltysik RC, Nelson RR, Frerichs RR, Noskin GA, Belknap SB, Bennett CL (2007). Discriminating inhalational anthrax from community-acquired pneumonia using chest radiograph findings and a clinical algorithm. *Chest*, 131, 489-495. DOI: 10.1378/chest.06-1687

[16]Green D, Hartwig D, Chen D, Soltysik RC, Yarnold PR (2003). Spinal cord injury risk assessment for thromboembolism (SPIRATE study). *American Journal of Physical and Medical Rehabilitation, 82*, 950-956. URL: http://journals.lww.com/ajpmr/Abstract/2003/12000/Spinal_Cord_Injury_Risk_Assessment_for.7.aspx

[17]Grobman WA, Terkildsen MF, Soltysik RC, Yarnold PR (2008). Predicting outcome after emergent cerclage using classification tree analysis. *American Journal of Perinatology, 25*, 443-448. DOI: 10.1055/s-0028-1083843

[18]Collinge W, Yarnold PR, Raskin E (1998). Use of mind/body self-healing practice predicts positive health transition in chronic fatigue syndrome: a controlled study. *Subtle Energies & Energy Medicine, 9*, 171-190. URL: http://journals.sfu.ca/seemj/index.php/seemj/article/view/256

[19]Soltysik RC, Yarnold PR (2010). Automated CTA software: Fundamental concepts and control commands. *Optimal Data Analysis, 1*, 144-160. URL: http://odajournal.com/2013/09/19/62/

[20]Yarnold PR, Soltysik RC (2014). Globally optimal statistical classification models, I: Binary class variable, one ordered attribute. *Optimal Data Analysis, 3*, 55-77. URL: http://optimalprediction.com/files/pdf/V3A17.pdf

[21]Yarnold PR, Bryant FB (2015). Obtaining a hierarchically optimal CTA model via UniODA software. *Optimal Data Analysis, 4*, 36-53. URL: http://optimalprediction.com/files/pdf/V4A11.pdf

[22]Yarnold PR, Bryant FB (2015). Obtaining an enumerated CTA model via automated CTA software. *Optimal Data Analysis, 4*, 54-61. URL: http://optimalprediction.com/files/pdf/V4A12.pdf

[23]Yarnold PR (2015). Distance from a theoretically ideal statistical classification model defined as the number of additional equivalent effects needed to obtain perfect classification for the sample. *Optimal Data Analysis, 4*, 81-86. URL: http://optimalprediction.com/files/pdf/V4A15.pdf

2155-0182/10/$3.00

[24]Yarnold PR (2015). Optimal statistical analysis involving multiple confounding variables. *Optimal Data Analysis, 4,* 107-112. URL: http://optimalprediction.com/files/pdf/V4A18.pdf

[25]Yarnold PR, Soltysik RC (2014). Globally optimal statistical classification models, II: Unrestricted class variable, two or more attributes. *Optimal Data Analysis*, *3,* 78-84. URL: http://optimalprediction.com/files/pdf/V3A18.pdf

[26]Yarnold PR (2013). Standards for reporting UniODA findings expanded to include ESP and all possible aggregated confusion tables. *Optimal Data Analysis*, *2,* 106-119. URL: http://optimalprediction.com/files/pdf/V2A19.pdf

Author Notes

The study analyzed de-identified 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