Optimal Data Analysis
Vol. 5 (October 8, 2016), 116-117
Copyright 2016 by Optimal Data Analysis, LLC
2155-0182/10/$3.00

ODA *vs.* Chi-Square: Differentiating Statistical and Ecological Significance

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

Statistical and ecological significance are rarely distinguished in reports employing chi-square analysis to test statistical hypotheses. This note elucidates this distinction, comparing chi-square with ODA using an application relating smoking knowledge and smoking behavior.[1]

Data are drawn from a study assessing the relationship of knowledge about effects of smoking (attribute) and smoking behavior (class variable) of high school boys and girls (stratification variable). Knowledge was a nominal indicator if an observation scored at or above versus below the sample mean (dummy-coded as 1 and 0, respectively) of a 66-item multiple-choice knowledge test on the effects of smoking. Subjects self-identified themselves as a "smoker" or "non-smoker" (dummy-coded as 1 and 0, respectively). And, for gender, boys and girls were dummy-coded as 1 and 0, respectively. Table 1 presents the data for this study.

Table 1: Smoking Study Data

	Boys		*Girls*	
Knowledge	Smoker	Non-Smoker	Smoker	Non-Smoker
≥Mean	37	45	26	75
<Mean	35	31	43	44

Chi-square analysis was used to assess if knowledge and behavior were independent, separately by gender. For girls a statistically significant chi-square statistic was obtained (chi-square=12.5, df=1, n=188, p<0.0004), but for boys the chi-square statistic was not statistically significant (chi-square=1.0, df=1, n=148, p< 0.33). The author concluded: "…the results of this study show a major difference between the sexes for the relationship in question" (p. 54).

ODA was used to analyze data in this design as described earlier: smoking behavior is the class variable and knowledge is the *ordered* nominal attribute (i.e., ≥mean is quantitatively greater than <mean), and analysis is conducted separately for boys and girls.[2,3] Non-directional (exploratory) analyses were conducted to "level the playing field" because chi-square analysis is inherently non-directional (i.e., "two-tailed"). For females the ODA model was: if knowledge score is less than the mean then predict that the observation is a smoker; otherwise predict a non-smoker: ESS=25.3 (weak-moderate effect), p<0.0007. For males the identical ODA model was obtained: ESS=7.8 (relatively weak effect), p<0.22. Thus, in terms of reliability the effect

Copyright 2016 by Optimal Data Analysis, LLC
2155-0182/10/$3.00

for females is highly statistically significant, but in terms of strength the effect barely surpasses the upper-bound criterion for a relatively weak effect (i.e., ESS<25).[2,3] Note that if the sample size in each cell (Table 1) is increased five times then, although ESS remains the same, $p<0.039$ for boys. If a directional *a priori* hypothesis is tested (e.g., greater knowledge predicts less smoking) then a four-fold increase in the sample size in each cell yields $p<0.034$.

Chi-square analysis used in the manner presented herein is ubiquitous in the literature, and it is important that researchers understand the distinction between statistical reliability (*p*) and ecological significance (ESS)—including the potential malleability of *p* versus stability of ESS, in interpreting and discussing findings. It is also imperative from a statistical validity perspective to ensure that underlying nominal variables are in fact categorical and not ordered, as was the case presently, if chi-square analysis is used because it assumes a qualitative variable.[3,4]

Finally, it is recommended that manual parsing of ordered or multicategorical attributes[5] or class variables[3] (in measurement instrument rating scales, or during statistical analysis) is avoided, and instead ODA or CTA are allowed to identify optimal thresholds and groupings (respectively) that maximize predictive accuracy and avoid paradoxical confounding.[2,3]

References

[1]Briney KL (1967). Relation of knowledge of effects of cigarette smoking to the practice of smoking among high school students. In: Zagona SV (Ed.), *Studies and issues in smoking behavior*. Tucson, AZ: University of Arizona Press (pp. 53-55).

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

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

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

[5]Yarnold PR (2014). "Breaking-up" an ordinal variable can reduce model classification accuracy. *Optimal Data Analysis*, *3*, 19.

Author Notes

The study analyzed de-individualized 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