

Detecting and Explicating Interactions in Categorical Data

By: William N. Dudley, Colleen DiIorio, and Johanna Soet

Dudley WN, DiIorio C, Soet J. (1999). Detecting and explicating interactions in categorical data. *Nursing Research*, 48, 53-56

Made available courtesy of LIPPINCOTT WILLIAMS & WILKINS:
<http://www.lww.com/nursing-research-journal/>

This format of the article is not the final published version.

*****Note: Figures may be missing from this format of the document**

Detecting and explicating interactions in categorical data analyses using cross tabulation and the $[\text{chi}]^2$ statistic can provide salient tests of hypotheses concerning the relationship between two variables measured at the nominal or ordinal levels. For example, researchers usually employ categorical analysis when they are interested in whether members of one group (e.g., males vs. females) differ in the proportion falling into two or more levels of a dependent variable (e.g., in favor of or opposed to sex education in public schools). In this case, the data can be expressed as a two-way table and hypotheses tested with the $[\text{chi}]^2$ statistic. Interpretation of this simplest of two-way tables is straightforward. However, research questions are often more complex than this simple example both in the number of predictor variables and the number of levels of each variable. Researchers typically include other predictor variables (e.g., race, academic status, marital status) to gain a better understanding of more complex relationships among predictors and outcomes. In addition, researchers often employ measures that have more than two levels (e.g., income, race, treatment type, academic status), and they often choose to combine levels in one or more variables to simplify the analyses, meet assumptions, or clarify the results.

The inclusion of more than one predictor leads to a model with multiway tables that can provide the researcher with thorny problems of analysis and interpretation. For example, suppose the previous model is expanded by including race (e.g., African American, Hispanic, and White). While the data could be analyzed using two separate two-way tables (one analysis for gender and one for race), this approach does not provide an opportunity to test for a possible race gender interaction. Here, interaction means that the effects of one predictor variable are not the same at all levels of a second predictor variable (Keppel, 1991). One possible race x gender interaction in this case might be that ethnic differences are only seen in male respondents.

There are several approaches to testing for possible interactions in categorical data. One solution is to use logistic regression with race, gender, and the race x gender interaction term as predictors of the dichotomous outcome variable (Hosmer & Lemeshow, 1989). A second alternative is to use log linear programming to test this interaction (Kennedy, 1992). Both methods are well-documented and commonly employed. However, both of these methods require a more extensive statistical background to conduct and interpret than does the traditional cross tabulation analysis. For example, the researcher may have to compute dummy variables or interaction terms. In addition, given a significant interaction term, steps must be completed to understand and

explicate the nature of the interaction. In logistic regression and log linear modeling this often requires plotting of proportions or logits across the levels of the other predictors (Hosmer & Lemeshow, 1989, p. 65; Kennedy, 1992, p. 205) which, while fairly easy to do with the 2×3 table discussed above, can become complicated with more variables and/or more levels per variable. The addition of one more predictor in this simple model more than doubles the possible combinations of main and interaction effects that may have to be specified in the model. Without question, logistic regression and log linear programming are useful analytic methods. However, the researcher who has initially chosen to employ crosstabs for simplicity of use and interpretation may be daunted by the technical expertise needed to implement and interpret the more complex analyses.

Chi square Automatic Interaction Detection

Solutions to the problems with interpretation of interaction effects have been sought for many years. In 1964 Sonquist and Morgan proposed a method for automatic interaction detection in complex parametric analysis. More recently, Kass (1980) proposed a Chi square Automatic Interaction Detection (CHAID) method for the detection of interactions in categorical data. This method has recently been operationalized in a menu-driven windows-based module by SPSS of the same name-SPSS CHAID (Magdison, 1993). CHAID is a data analysis tool that can be very useful both in the interpretation of interaction effects and in making decisions about how to combine categories within a variable to arrive at the most parsimonious model. The menu-driven program is easy to use, and most important, the results can be presented in an easy-to-understand graphical form.

CHAID has thus far been employed largely by analysts in applied marketing who specialize in extracting information from existing data sets. For example, Dun and Bradstreet (1995) report using CHAID as a data mining technique. Even a casual search of the Internet will locate many data mining sites similar to the Dun and Bradstreet site; however, the scientific literature reporting CHAID is far more sparse. Although rare, there are a handful of publications by researchers who employ CHAID in marketing (Baron & Phillips, 1994), medicine (Hill, Delaney, & Roncal, 1997), and nursing (Hogan & Smith, 1987). The paucity of papers using CHAID indicates that these methods are largely untapped by nurse researchers and others in the health sciences. This paper is intended to bring to the nurse researcher's attention a statistical technique that may be useful in the analysis of complex patterns of relationships among several categorical or ordinal variables.

The Process

CHAID is a statistical technique designed to partition a sample into mutually exclusive groups and subgroups based on a single categorical or ordinal outcome and several categorical or ordinal predictors. The analysis proceeds in a stepwise fashion in which the most significant predictor is used to partition the entire sample into two or more mutually exclusive subgroups. For example, suppose that in the first step, gender was the most significant (having the largest $[\text{chi}]^2$) predictor of sex education attitudes. This partitioning would result in two subgroups: males and females. In the second step, the analysis is performed on each of these subgroups in turn. As in the first step, the cases in a given subgroup are partitioned by the next most significant predictor of the original outcome (sex-education attitudes). The next most significant predictor variable may differ from subgroup to subgroup. For example, in the second step, race may be the most significant

predictor for males and marital status may be the most significant predictor for females. In this case, males would be partitioned by race-females by marital status. Once the subgroups at a given level have been partitioned, the process begins again with all newly partitioned subgroups analyzed in the identical stepwise process. The analysis continues in this manner until there are no more significant predictors, or until some user-specified constraint has been met.

Several important aspects of the process should be noted. First, given the large number of comparisons that may be performed, one may choose to employ a Bonferroni adjustment to control for Type I error. Second, the analyst may choose to allow the program to collapse across categories of a given predictor variable based on level of significance, size of the subgroups, level of data (ordinal versus nominal), etc. Third, if the researcher has a theoretically based rationale for testing variables in a particular order, he or she may force the order of entry rather than rely on the automated stepwise process. Finally, test statistics are available for inspection and reporting at each level and the final results of the analysis are also presented in a tree diagram that allows for interpretation of very complex interactions.

An Example

To illustrate the utility of CHAID, a mock study on attitudes toward sex education was constructed. The data were designed to highlight the detection of complex interactions, and the process of collapsing two or more categories into one. In this data set, complex interactions will be seen that would not likely come to the researcher's attention with traditional crosstab analysis and would be difficult to interpret in logistic regression or log-linear modeling.

The analysis process employed by CHAID can perhaps best be illustrated by returning to traditional crosstab analysis. Recall that in traditional analysis, the sample is partitioned into groups based upon the dependent and predictor variables. For example, in Table 1 we have reproduced a simple 2×2 table used to test for gender differences in attitudes toward sex education. As the table shows, males are far more likely to approve of sex education than are females ($[\chi]^2(1) = 14.80, p < 0.001$). A second analysis (Table 2) depicts ethnic differences in approval of sex education. While the results presented in this table are also significant at the 0.001 level ($[\chi]^2(1) = 13.96, p < 0.001$), the differences are not as easy to tease apart as in the previous example. It seems as though White and Hispanic respondents may not differ significantly. One could run pair wise analyses (while controlling for Type I error) or use standardized residual to establish where the differences lie in these data. However, the conclusions could be invalid if predictors outside the model interact with those within the model (see Table 3). In this example, the race differences are striking for males but not for females.

Gender	Favor Sex Education		Total [n (%)]
	Yes [n (%)]	No. [n (%)]	
Male	31 (41.3)	44 (58.7)	75 (50)
Female	10 (13.3)	65 (86.7)	75 (50)
Total	41 (27.3)	109 (72.7)	150 (100)

Note. $\chi^2 (1) = 14.80, p < .001$.

TABLE 1. Cross Tabulation of Gender With Attitude Toward Sex Education

Race	Favor Sex Education		Total [n (%)]
	Yes [n (%)]	No [n (%)]	
White	11 (22.0)	39 (78)	50 (33.3)
Black	23 (46.0)	27 (54.0)	50 (33.3)
Hispanic	7 (14.0)	43 (86.0)	50 (33.3)
Total	41 (27.3)	109 (72.7)	150 (100)

Note. $\chi^2 (1) = 14.80, p < .001$.

TABLE 2. Cross Tabulation of Race With Attitude Toward Sex Education

Gender	Race	Favor Sex Education		Total [n (%)]
		Yes [n (%)]	No [n (%)]	
Male	White	5 (20)	20 (80)	25 (16.6)
	Black	21 (84)	4 (16)	25 (16.6)
	Hispanic	5 (20)	20 (80)	25 (16.6)
Female	White	6 (24)	19 (76)	25 (16.6)
	Black	2 (8)	23 (92)	25 (16.6)
	Hispanic	2 (8)	23 (92)	25 (16.6)
Total		41 (27.33)	109 (72.7)	150 (100)

TABLE 3. Multiway Table for Condom Use With Race and Gender

The interaction with the original data plus a third predictor (marital status) was calculated (see Figure 1) using CHAID. In the CHAID output, each box is comparable to a cell in a typical cross

tabulation (such as seen in Tables 1-3). Below each cell, one finds the name of the dimension along which the next partition is based. For example, the most salient dimension for differences in sex education in this data set is gender. Note that this figure reflects the dummy coding scheme used ("1" represents opposition to sex education). In addition, note that the terminal nodes are numbered consecutively immediately below the cell.

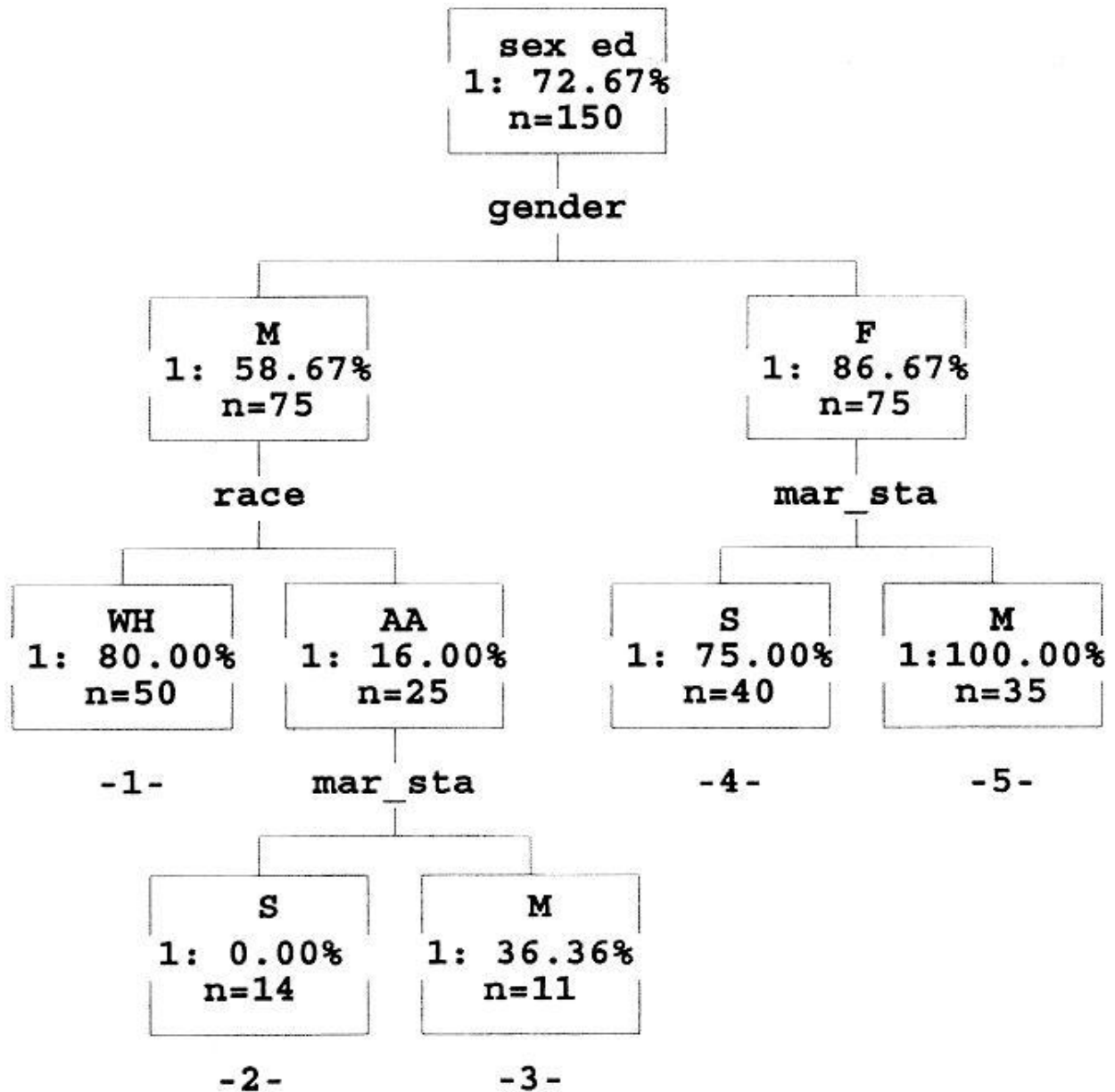


Figure 1. CHAID output for condom use by race, gender, and marital status.

Figure 1 shows that the sample is partitioned into males and females just as in Table 1. While 86.67% of females are opposed to sex education only 58.7% of males are opposed. Males are also partitioned by race. The program collapsed racial groups so that for males, Whites and Hispanics are not significantly different in attitudes toward sex education (80% are opposed) but together they are significantly different from African Americans males, of whom only 16% are

opposed. For females, the most significant predictor at this level is marital status. Of the 40 single females, 75% are opposed to sex education, while 100% of the 35 married women are opposed. Notice that at this point, two two-way interactions are detected: a gender x race interaction and a gender x marital status interaction. At the third level, marital status is a significant factor for African Americans; of the 14 single African American males, none are opposed to sex education while 36.36% of the married African American males are opposed. The analysis stops at the level of marital status for women.

This very complex result was based on data generated to illustrate how the CHAID analysis takes place. However, in practice results just as complex have been encountered that would probably not have come to light or have been easily explained using more traditional analysis. Note that the analysis reported here was done with no restriction placed on order of entry. If the primary focus of the research had been to investigate ethnic differences, race may have been forced into the model first; the results would have differed, as would the conclusions.

Uses for CHAID

CHAID may be used as the principle hypothesis testing analysis when dealing with sets of categorical predictors and categorical outcomes or as an exploratory technique for the discovery of possibly important interactions that may be difficult to detect or understand with more traditional procedures. A more refined analysis using logistic regression might evolve from these preliminary results. CHAID analysis can locate nonlinear relationships and, thus, can be used to diagnose violations of assumptions, in a logistical regression that is not performing as expected. CHAID might also be used in applied settings. It seems reasonable that segmentation of a target population could be a useful strategy for health educators. For example, a sex education program might be tailored to subgroups identified in a CHAID analysis. Health educators have begun to adopt social marketing strategies such as market segmentation to target specific populations to affect health behavior change (Andreasen, 1995). CHAID may be a useful tool for the health educator in making the most efficient use of limited resources and in developing the most effective messages for a targeted population.

Limitations

The use of automated empirically based procedures carries a risk of finding statistically significant results that may make no sense theoretically. Just as with stepwise regression, the researcher must exercise judgment and not rely on the software to make decisions crucial to the integrity of the research (Pedhazur, 1982). CHAID allows the researcher to exercise control (e.g., in collapsing categories, forcing order of entry, and limiting the levels of partitioning). The researcher is strongly recommended to use these options. A second limitation is that this program is designed for data that are categorical or ordinal in nature. If one or more of the predictors are continuous in nature, and there are no well-established groupings, other methods such as logistic regression may be more appropriate. A third limitation is that with repeated segmentation of the sample, cell sizes can become very small after only a few iterations. Thus, complex interactions that we have depicted may only be practical with large samples. Another concern is that the CHAID methodology is a "forward only" stepwise process. Once a segmentation has taken place, it is fixed and will not be reevaluated. This is very different from more traditional stepwise regression analysis in which variables can come in and out of the model. Finally, because CHAID has thus far been used primarily for data mining and market analysis, researchers in

other fields who use CHAID may have to devote some energy to explaining to the reader how the process works.

Conclusions

The analysis and interpretation of categorical data can be far more demanding than one might expect. Of particular difficulty is first detecting and then explaining the nature of complex interactions among two or more predictors. Although logistic regression and log linear programming may be used in these analyses, interpretation of the results can be problematic. CHAID is one tool that the researcher can employ with relative ease to gain a clear understanding of complex relationships among categorical variables. The programming is menu driven and uses SPSS data files. The graphical output is intuitive and especially clear and useful in explaining complex interrelationships among several variables. Those researchers who deal with categorical data may find this rather new technique useful and easy to implement.

References

- Andreasen, A. R. (1995). *Marketing social change: Changing behavior to promote health, social development, and the environment*. San Francisco: Jossey-Bass.
- Baron, S., & Phillips, D. (1994). Attitude survey data reduction using CHAID: An example in shopping centre market research. *Journal of Marketing Management*, 10, 75-88.
- Dun & Bradstreet. (1995, September). An overview of data mining at Dun & Bradstreet. Cambridge, MA: Data Intelligence Group [On-line]. Available: www.santafe.edu/~kurt/text/wp9501/wp9501.shtml
- Hill, D. A., Delaney, L. M., & Roncal, S. (1997). A chi-square automatic interaction detection (CHAID) analysis of factors determining trauma outcomes. *Journal of Trauma*, 42(1), 62-66.
- Hogan, A. J., & Smith, D. W. (1987). Patient classification and resource allocation in Veterans Administration nursing homes. *ANS: Advances in Nursing Science*, 9(3), 56-71.
- Hosmer, D. W., Jr., & Lemeshow, S. (1989). *Applied logistic regression*. New York: Wiley.
- Kass, G. V. (1980). An exploratory technique for investigating large quantities of categorical data. *Journal of Applied Statistics*, 29(2), 119-127.
- Kennedy, J. J. (1992). *Analyzing qualitative data: Log-linear analysis for behavioral research* (2nd ed.). New York: Praeger.
- Keppel, G. (1991). *Design and analysis: A researcher's handbook*. (3rd ed.). Englewood Cliffs, NJ: Prentice Hall.
- Magdison, J. (1993). *SPSS for Windows CHAID, Release 6.0*. Chicago: SPSS.
- Pedhazur, E. J. (1982). *Multiple regression in behavioral research*. (2nd ed.). New York: Holt, Rinehart & Winston.
- Sonquist, J. N., & Morgan, J. A. (1964). *The detection of interaction effects* (Monograph #35). Ann Arbor, MI: Survey Research Center, Institute for Social Research, University of Michigan.