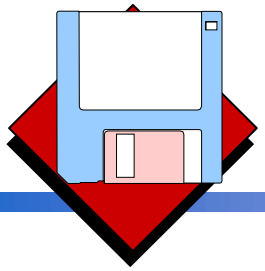

Data Use

Some specialized problem-solving applications of CHAID

By Steven Struhl

Editor's note: Steven Struhl is senior vice president, senior methodologist at Harris Interactive

CHAID and its related techniques (especially CHAID/ CART) can solve some thorny data analysis problems that will defeat most other methods. As a reminder, CHAID works by splitting and re-splitting a sample into groups that differ as much as possible on some dependent variable. If your dependent variable was purchase intent, then when the analysis concluded, some group(s) would have relatively high purchase intent and some relatively low intent.

CHAID allows you to look at all the independent variables that would produce significantly different groups (based on whatever level of significance you choose), and select one as the basis for splitting the sample. Once you have done this, the procedure will examine each group split off, and find the best differentiators again. It will keep re-splitting until no more significant variables can be found, or until you reach a lower size limit on the groups split off.

A powerful tool: free combination or optimal recoding

One of this procedure's most powerful features is its ability to allow independent or predictor variables to combine freely, in order to produce the most dramatic differences between groups formed. Suppose you have five regions of the country (N, S, E, W, and S, coded 1, 2, 3, 4, 5). CHAID will combine these in any way that gives the strongest between-group differences on your dependent variable. You might, for instance, find regions 1, 3 and 4 in one group, and regions 2 and 5 in the other. This feature sometimes is called *optimal recoding*, because you can develop new coding schemes for your independent variables using it.

I prefer to think of optimal recoding as something like super-intelligent cross-tabs, that can find the best way to combine columns of data in the most strongly contrasting way. The special applications here rely on CHAID's ability to perform optimal recoding. The readily available analysis packages that run on PCs will do this, although some may be more flexible than others.

SPECIAL APPLICATIONS

1. Surely there's some relation ...

Sometimes you will encounter a situation where you need to prove (or to disprove) definitively that some relationship exists between two variables. Suppose you have a client (in our example, the second assistant brand manager) who is certain that a certain favorite feature (in our example, discount coupons) will have an impact on purchase intent. You do cross-tabs and nothing emerges directly. Yet the second assistant is still sure there's some interaction hidden there.

"Optimal recoding" will allow you to get a definitive answer, even if purchase intent is scaled 0-100 as in the example.

The steps in this procedure are as follows:

- Set the acceptable significance (α) to 0.99. That is, let any variable with more than a 1% chance of being significant appear on the list of possible predictors.

Reprinted from *Quirk's Marketing Research Review*

- If you have a continuous independent variable, divide it into numerous ranges (e.g., 15, 20, or more).
- Allow the variable of interest to enter, and observe the significance value of "optimal" split.

This will provide far better estimates than trying many alternative splits "by hand," since CHAID and CHAID/ CART programs will keep track of how many alternative splits you are examining in any comparison and adjust significance levels accordingly. If you do not make an adjustment for the number of alternative ways you can split the data, you likely will find spuriously significant ones—leading to the unfortunately familiar wild goose chase (at the least). If you try to make the adjustment yourself, you will probably take too conservative an approach. All up-to-date CHAID programs have rules built in that handle this problem nicely.

In this example, much to the second assistant brand manager's dismay, the coupons showed no significant interaction with purchase intent. The utilities come from a full profile conjoint task in which coupons were one of the attributes tested. Those who liked the coupons more (had higher utilities for them) did not show higher purchase intent.

2. Effects of all possible patterns of usage

You may encounter a problem where you need to know how usage patterns in the product category interact with liking or usage of your brand. For instance, suppose your client, the Cache-X® Charge Card, wants to know how ownership of various other cards, singly and in combination, interact with the dollar amount charged on their card.

If you do not have too many other competitors, CHAID/ CART can help you answer this question. Here is one way you could do this:

- Create a "meta-variable" showing all products used. You can do this simply by recording usage in consecutive columns, and then calling all the columns together one variable for instance, for 5 products, you would define a position variable, with 1 in the first position meaning owns product 1, zero in first position meaning does not own product 1, and so on; in this case, 01101 would mean: doesn't own product 1 and product 4, owns products 2, 3 and 5.
- SPSS, for instance, will allow you to say that usage of the Cache-X card is in column 4, Card 2 is in column 5, and so on, and then in addition say columns 4 to X contain one variable five columns long, called (for instance) "USAGE."
- Use this meta-variable as the predictor, allowing it to combine freely, to create "optimal recodings."

This "optimal" recoding takes some study, but shows that:

- Those with 4 or more cards charge significantly less than all others on the Cache-X card;
- Those with Cache-X and card 5 charge significantly more than all others;
- Those having Cache-X and no card 5 charge a moderate amount relative to all others.

With the more flexible programs, the limits on products entered is your personal limit for reading and interpreting output. Since everybody in our example used the Cache-X Card, we had only 16 combinations with four other brands. Results would get much harder to read if you had 5 products all of which could appear or not (32 possible combinations). Several programs can handle 6 products (or even 7) with 2 usage levels (that is, 64 or 128 possible combinations of usages). The question is whether you will find anything you can interpret among so many combinations.

If you need to look at 3 usage levels (for instance, non-users vs. light users vs. heavy users) that effectively limits you to 4 products or fewer in one analysis (with 4 products at 3 usage levels, you would have 81 possible combinations).

You can, of course, break a problem involving 3 usage levels into several separate two-level problems: one contrasting combinations of non-users and heavy users, one contrasting combinations of light users and heavy users, one contrasting combinations of non-users and light users. This would give you three analysis, and

with 4 brands each would involve 16 combinations.

Another highly useful characteristic of CHAID and CHAID/ CART emerges in this type of analysis. If nobody uses a certain combination of brands, that combination simply does not appear in your analysis. Unlike many analysis-of-variance based procedures, CHAID does not get thrown by "empty cells." You can have all or just a few of all possible combinations and the analysis will still provide accurate results.

3. Surely there's some relation... (Part 2)

Suppose the second assistant brand manager returns to you. He (or she, if this can be determined) just knows you didn't do a thorough enough job. "What if people already like another feature, like the adult videos, or the great books set?" s/he complains. "If people already liked one of those, then if they like the coupon, will they be more interested in buying?"

Fortunately, CHAID-type procedures let you answer questions like these definitively. That is, you can see if the feature adds incremental value where people already like other features. In somewhat more formal terms, you can see if an effect appears, conditional upon other effects.

Here is one possible method:

- Allow all significant variables to enter the analysis
- After the tree is complete, see if the variable of interest (in our example, the coupons) enters at the (a) = 0.99 level or better. (Again, you set the significance so that all variables with at least a 1% of being significant can get onto the list of candidate variables.)
- A for a more exhaustive method: try to "force in" the variable at every spot along the way.

In this instance, the second assistant still had no luck. The coupons got "forced into the analysis" after all significant predictors emerged, and still nothing happened.

To understand the situation fully, consider that the product is a lifetime subscription to *Tall Tales Magazine*. Respondents in the study were told they could get this subscription at various prices, with a variety of other special bonus products included or not. Using a full profile conjoint design, 16 alternative "bundles" of special products were developed. Respondents gave their overall purchase intent rating to each of these on a 0 to 100 scale, and from these we developed utilities for each add-on. In this actual (disguised) example, purchase intent (at the insistence of the first assistant) was recoded to a 3-category variable, with different points on the 0—100 scale used as cut-offs for "mid" and "high" level s of purchase intent.

This was an arbitrary procedure and might possibly have compromised the analysis. Fortunately, when all others were safely sleeping, we went back and rechecked with the target variable allowed to act as a scaled variable.

Conclusions were the same. However, this did illustrate an additional point of flexibility for CHAID—it is one of a very few procedures that can legitimately handle a dependent variable that is categorical or continuous.

To conclude

You can use the optimal recoding abilities of CHAID and CHAID/CART in many other ways, limited only by your imagination. They are particularly powerful at finding non linear patterns of interactions in variables with many categories, or in continuous variables. Added to this these procedures can investigate interactions of these types conditionally--as in the example directly above. Even better, you can do all this without building an explicit model, as in analysis of-variance based procedures, and without worrying about "empty cells. " The second assistant brand manager might not like the findings, but these features make these procedures distinctive and powerful additions to your armamentarium of data analysis methods.