Grow a Tree to Get your Fruits: A Guide to Using and Understanding CHAID Analysis

By Ruomei Feng *(Marketing Scientist)*

Do you want to attract more customers? Are you wanting to increase the response rate to your mass-mailed campaign? Have you ever been puzzled by who stopped buying your products? All these business issues are more-or-less related to "who are they" questions. CHAID analysis is a useful tool to help figure out "who they are."

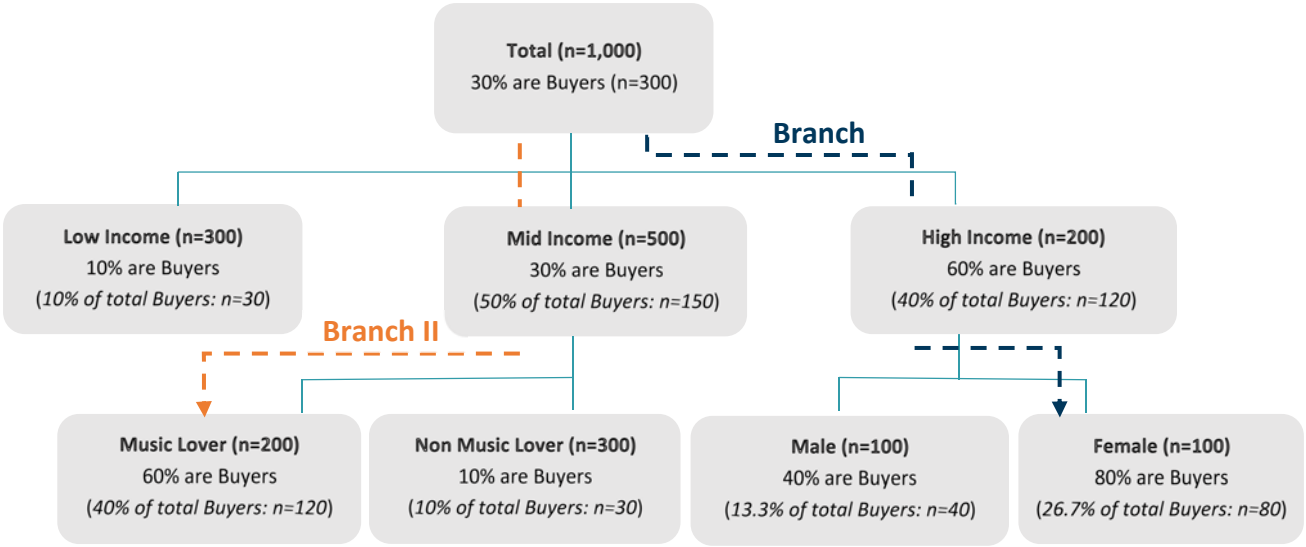
Understanding CHAID

CHAID stands for Chi-square Automatic Interaction Detection. By name, this method uses Pearson's Chi-square test of independence to detect interaction between variables (*predictors*). The predictors define your target, such as who has higher potential to be your new customers, who tends to respond to your fliers, or who is at risk of leaving your business. CHAID analysis is also known as a decision tree. Therefore, more specifically, it grows a "tree," where the branches bring us to our target by identifying meaningful characteristics about groups of respondents.

The major advantages of CHAID include:

- Highly visual output
- An absence of equations
- Very easy to understand results
- Clear implications

Here is an example: We would love to know who tends to purchase products in our category. Are they male or female? Their ages? What income levels are they at? Do they have any hobbies? We explored the available information and grew a tree as follows.



In this example, among the population (*sample size n=1,000*), there are 300 category buyers.

- Taking **Branch I**, from high income to female, we can identify a group accounting for only 10% (*n=100*) of the population, and yet 80% of them are category buyers

- Taking **Branch II**, from mid income to music lover, we can find another group representing 20% (*n=200*) of the population, in which 60% of them are category buyers

Say we want to reach 200 category buyers and we know there are 300 category buyers in a population of 1,000. Without the tree, in order to reach 200 category buyers, we need to randomly contact 667 people (*67% of the population*).

But by targeting based on both of these branches, to locate 200 category buyers, we only need to access 30% of the population (***Branch I*** *and* ***Branch II*** *combined*) (*n=80 from* ***Branch I*** *and n=120 from* ***Branch II***). The efficiency in reaching buyers has essentially been doubled by growing and taking advantage of this tree, from 67% to 30%.

What Types of Variables You Need for CHAID

CHAID uses a chi-square distribution to find out what variables best separate the target. The implication is that we should use categorical variables to define our sub-groups of interest. We call these target variables.

Target variables show the classification of the groups that a tree will distinguish, such as:

- Buyer vs. Non-Buyer
- Satisfied vs. Dissatisfied
- Large Volume Buyer vs. Limited Volume Buyer
- Responded to Direct Mail vs. Didn't Respond

Often, we want to have a better understanding of the groups that are more valuable to the company. For example, we are more interested to determine who product buyers are, large volume buyers are, or those who are more satisfied with our services. But sometimes, we also want to target those with lower values, such as the customers we have lost, to reveal who may need special attention.

It is possible to expand to a three-level, or a four-level target variable, if necessary. However, in market research practice, it is not recommended to go beyond four levels for a target variable. While it is not a problem mathematically, the outputs are generally messy and difficult to understand. Simplification of the variables yields crisper and more actionable results for practical business application.

A benefit of CHAID analysis is it accommodates a large number (*a couple hundred*) of **predictors**. Predictors can be categorical or continuous variable types: The categorical predictors can be either two-level or multi-level. The analysis will divide continuous predictors (*e.g., age, number of loyalty programs, monthly mortgage/rent payment*) into buckets. The program decides where to cut and how many cutting points are needed in order to best differentiate the target from the non-target respondents. To make the cut points more useful or intuitive, the team may recode a continuous predictor into a categorical variable before the analysis is run and it will be treated as categorical.

When to use CHAID

CHAID analysis is an exploratory method with wide application. It fits in most types of market research studies: new product development, segmentation, satisfaction trackers, message trackers, ad efficiency, pricing, CRM, etc. As long as the goal is to discover more efficient ways to access and understand target group(s), CHAID analysis is a valuable technique to consider.

About the Author

Ruomei Feng has amassed ten years of experience in primary market research, and works across KS&R's teams on a broad array of complex projects. Her business analytics background includes modeling & trend analysis, data mining with large datasets, database/direct marketing, and efficiency measurement to provide business understanding and solutions.

Return To *Thought Leadership >*

KS&R Data to Knowledge