

TUTORIAL 1: BEGINNING A CHAID ANALYSIS

In this Tutorial we illustrate the basic functions and uses of SI-CHAID. We will show how to set up an analysis (.chd) file and grow a CHAID tree by using the standard CHAID algorithm, which is designed for a dichotomous or nominal dependent variable. In our example, we show how to determine CHAID segments that differ on response rates, and how gains charts can be used to predict the expected response from mailing/ targeting the most responsive segments. Tutorial #2 illustrates the use of the ordinal algorithm in SI-CHAID to identify segments best upon a profitability criterion. Both tutorials follow the analyses described in Magidson (1993).

The Data

In this tutorial, we will be using the SPSS file *subscrib.sav*, which contains information about a direct marketing promotion for a magazine subscription. Based on their response to this promotion, households were categorized as paid responders, unpaid responders, or nonresponders. Paid responders were households that returned a mail form, checked off the item that they would like to subscribe to the magazine, and later paid for the subscription. Unpaid responders were households that returned the form and checked off the item that they would like to subscribe to the magazine, but then cancelled their subscriptions prior to paying. Nonresponders includes all others (that is, households that did not request a subscription).

	age	gender	kids	income	bankcard	hhsz	occup	resp3	resp2	freq
1	1	1	1	2	7	2	1	4	1	1
2	1	1	2	4	2	1	4	1	1	1
3	1	1	2	3	2	1	4	1	1	1
4	1	1	2	3	1	1	4	1	1	1
5	1	2	2	4	2	2	3	1	1	1
6	1	2	2	5	2	1	4	1	1	1
7	2	1	1	7	1	4	2	1	1	1
8	2	1	1	6	2	5	3	1	1	1
9	2	1	1	4	2	5	2	1	1	1
10	2	1	1	1	2	3	3	1	1	1
11	2	1	2	8	2	2	1	1	1	2
12	2	1	2	8	1	2	1	1	1	1
13	2	1	2	7	2	2	4	1	1	1
14	2	1	2	7	1	4	4	1	1	1
15	2	1	2	6	2	2	4	1	1	1
16	2	1	2	4	2	2	3	1	1	1
17	2	1	2	3	2	2	3	1	1	1

The variables included in the file are:

AGE	age of head of household
GENDER	sex of head of household
KIDS	presence of children
INCOME	household income
BANKCARD	presence of bankcard
HHSIZE	household size

OCCUP	occupational status of head of household
RESP3	coded 1 for paid, 2 for unpaid responders and 3 for nonresponders.
RESP2	coded 1 for (paid and unpaid) responders, and 2 for nonresponders – to be used as the dependent variable in this tutorial
FREQ	number of cases (designated as a case weight in SPSS)

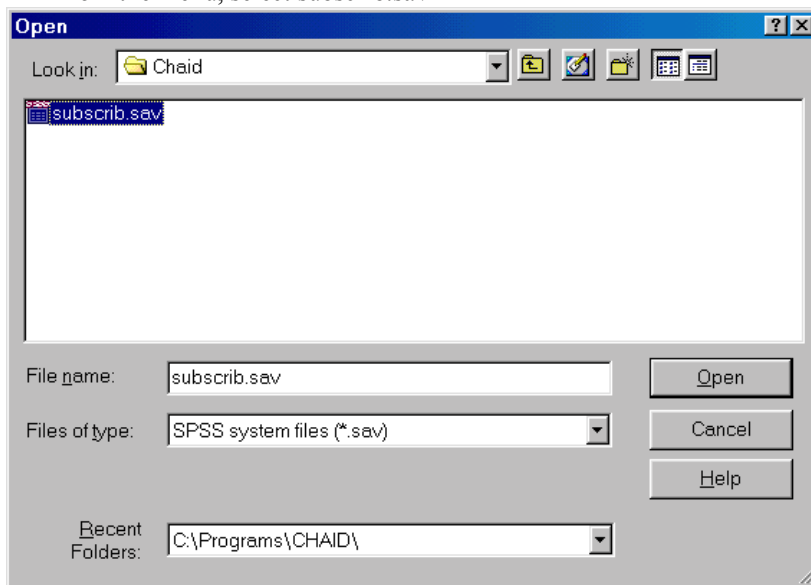
The purpose of our initial analysis is to identify household segments that are more likely to respond than other segments.

Setting up the Model

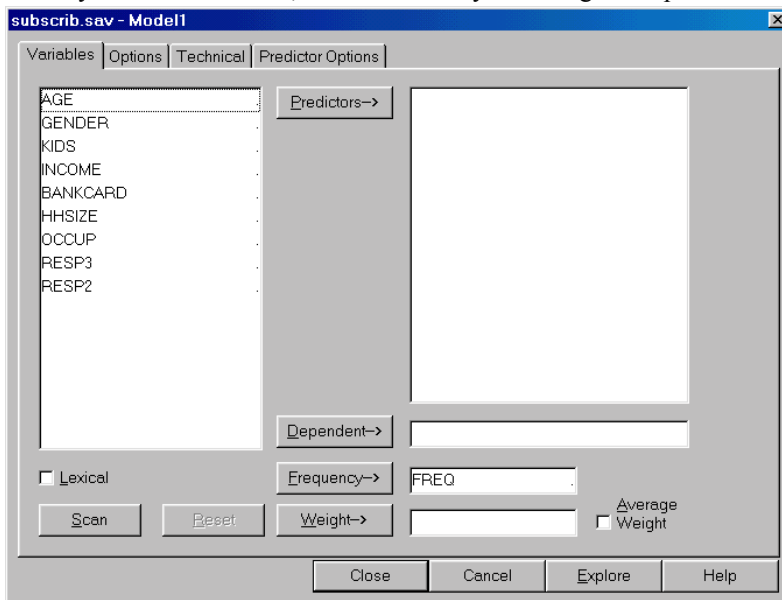
Opening the Data File

To open the file,

- Open ChaidDefine.exe from the CHAID Directory
- Go to the File Menu and click New
- From the menu, select subscrib.sav



Once you click on the file, the Model Analysis Dialog Box opens. It looks like this:



The Variables included the data file subscrib.sav are listed in the Assign Variables box on the left. (Notice that SI-CHAID automatically entered the variable FREQ in the frequency box because it was specified within SPSS to be used as a case weight when creating the SPSS save file.)

Assigning Variables

To begin a CHAID analysis, we need to select a dependent variable and at least one predictor. Optionally, one of two weight variables can be specified - a case weight (frequency) and a sampling weight (weight).

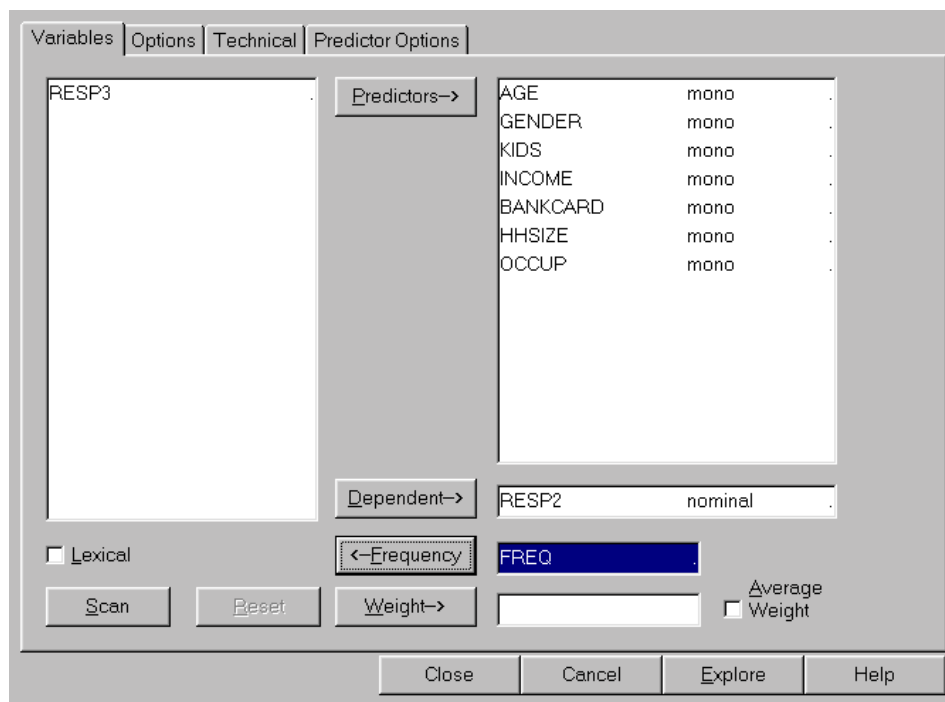
For this analysis, the dichotomous variable RESP2 will be the dependent variable. To select the dependent variable:

- Click on RESP2 in the Variables Box.
- Click on "Dependent" to move RESP2 to the Dependent Variable Box

Next, we will select the predictor variables. The predictor variables for this analysis will be AGE, GENDER, KIDS, INCOME, BANKCARD, HHSIZE, and OCCUP.

- Highlight AGE, GENDER, KIDS, INCOME, BANKCARD, HHSIZE, and OCCUP.
- Click on "Predictors" to move the above variables to the Predictor Variable Box.

The completed Model Analysis Dialog Box should look like this:



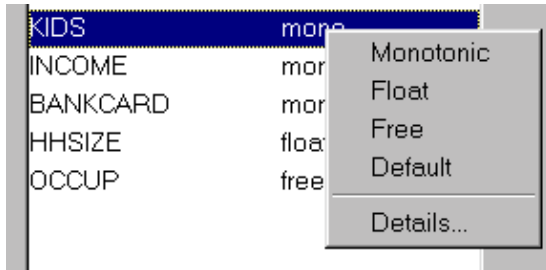
Scanning the Data

Now that you have set your analysis options, you are ready to scan the data file. To scan the file,

- Click on the Variables Tab
- Click on Scan Data

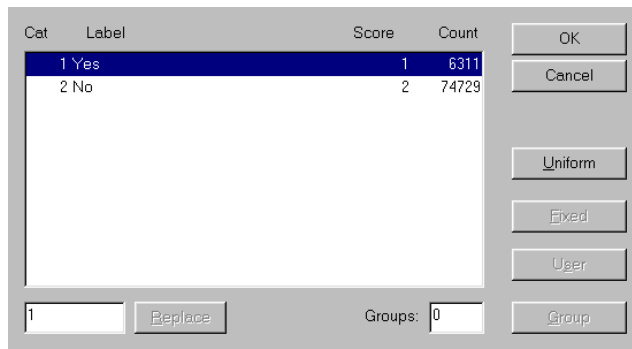
After the data scans, the default combine types appear next to each predictor. The combine type specifies how the categories of the predictor are allowed to merge. You can change the combine type for a predictor

from the Predictor Options tab or by right-clicking on the variable and selecting the desired combine type name from the pop-up menu.



- Right-click on OCCUP and select "Free" to define OCCUP as a free variable

You may view category labels by selecting *Details...* from this menu or by double-clicking on a predictor or the dependent variable name. This action brings up the category-labels window.

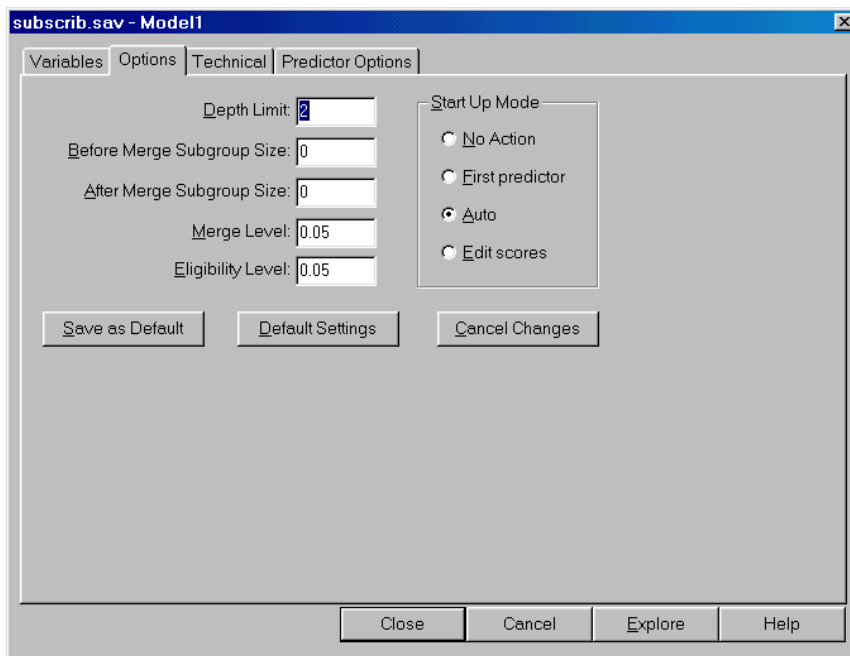


Setting Options

The Options Tab controls the operation of the CHAID segmentation algorithm, including the stopping rule and the minimum segment size.

- Click on the Options Tab to open the Options Dialog Box
- Double-click on the Depth Limit text box and enter 2 to set the analysis depth limit at 2. That tells SI-CHAID that the tree should expand to no more than two levels deep.
- Leave the other options, Merge Level and Eligibility Level, at their default levels.
- Select Auto in the Startup Mode Menu on the right. This tells SI-CHAID to run the analysis automatically.

Your Options Dialog Box should now look like this:

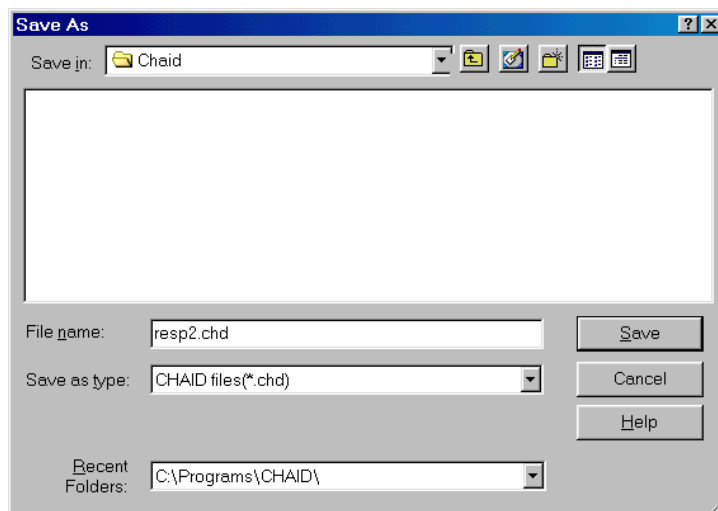


Growing a Tree

After you have set all the options, you are now ready to grow a segmentation tree.

➤ Click Explore

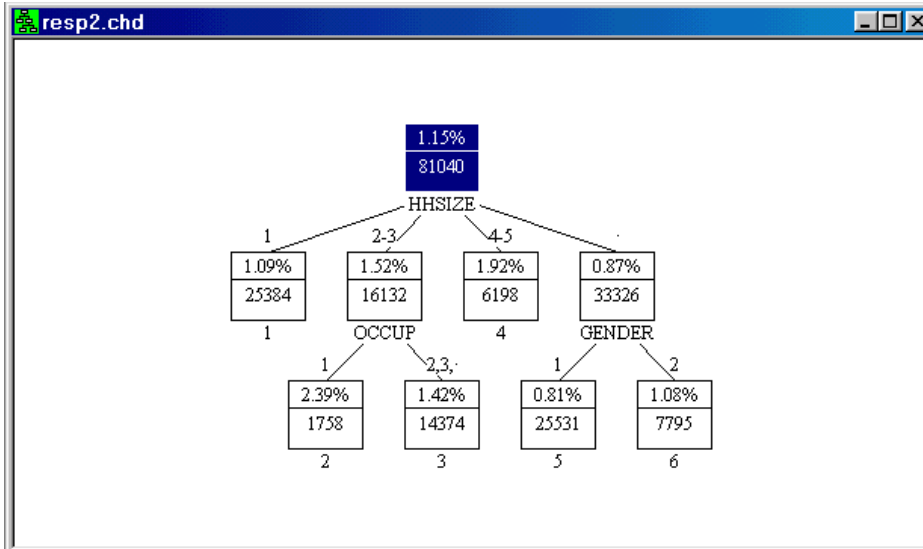
SI-CHAID automatically prompts you to save the new model with a Save As dialog box.



In the File Name box, type *resp2* to override the suggested filename and click on Save. That tells SI-CHAID to save your analysis settings to an analysis file with the name *resp2.chd*. All printed and saved output will be prefixed by the name *resp2*.

Growing a Tree in Automatic Mode

After you click Save, SI-CHAID automatically opens the ChaidExplore program and grows the tree.

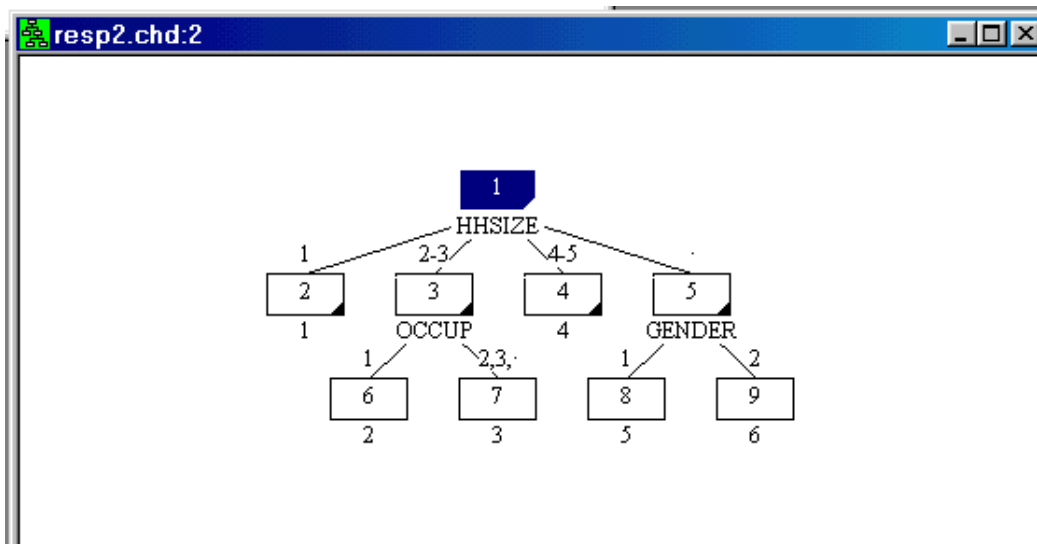


By default, SI-CHAID displays the tree diagram in local mode. The local mode displays detailed results within each node, and numbers each terminal node. The results of the CHAID tree shows 6 segments, details for which are displayed in each of the 6 terminal nodes. The highest response rate is obtained from segment 2, defined as households of size 2 or 3 (HHSIZE = 2-3) and occupation = 'white collar' (OCCUP = 1). Terminal node #2 shows that there are a total of 1,758 cases in this segment and the response rate is 2.39%. The next best segment is obtained from households containing 4 or more persons (terminal node #4), and the response rate for this segment is 1.92%.

For large trees, all terminal nodes may not be visible at once. In this case, a global tree view is useful to get a better feel for the entire tree. To switch to global mode,

- Click on Window
- Select New Tree Map

The Global Tree Window then appears



Gains Charts

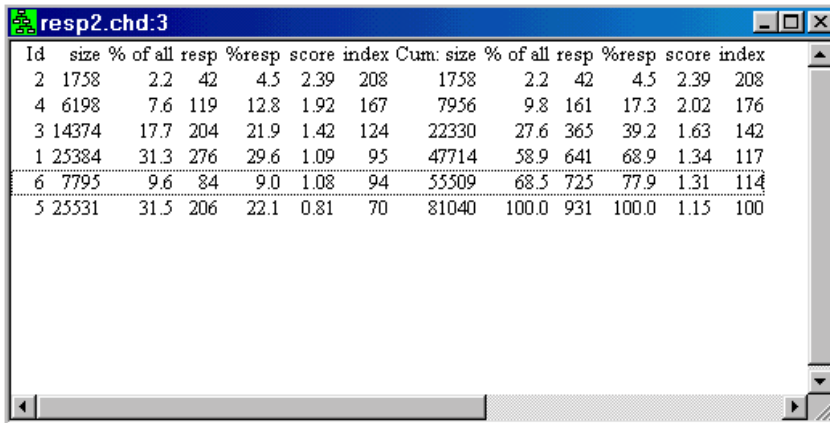
The results of a CHAID analysis can also be displayed in the form of **Gains Charts**, which sorts the segments from *best* to *worst* and also provides cumulative results expected if the best K segments (or best quantile of the respondents) are mailed. In our current analysis, *best* is defined based on the percentage of cases in the first category of the dependent variable (response rate).

Detailed Gains Charts

To produce a detailed gains chart corresponding to the current CHAID tree:

- Click on Window
- Select New Gains

SI-CHAID displays a detailed gains chart, where the segments are listed from best to worst.



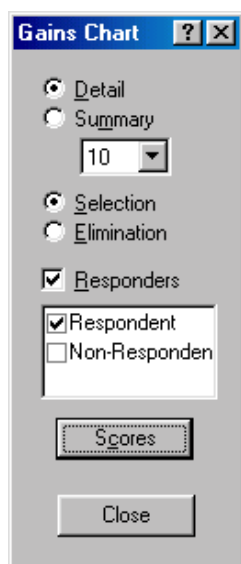
Id	size	% of all	resp	%resp	score	index	Cum: size	% of all	resp	%resp	score	index
2	1758	2.2	42	4.5	2.39	208	1758	2.2	42	4.5	2.39	208
4	6198	7.6	119	12.8	1.92	167	7956	9.8	161	17.3	2.02	176
3	14374	17.7	204	21.9	1.42	124	22330	27.6	365	39.2	1.63	142
1	25384	31.3	276	29.6	1.09	95	47714	58.9	641	68.9	1.34	117
6	7795	9.6	84	9.0	1.08	94	55509	68.5	725	77.9	1.31	114
5	25531	31.5	206	22.1	0.81	70	81040	100.0	931	100.0	1.15	100

The column labeled *Id* contains segment numbers. The next column (*size*) contains the number of cases in this segment, followed by a re-expression of segment size in terms of a percentage (% of all). The 4th column (*resp*) contains the number of responders in the segment, followed by a re-expression of this quantity in terms of percentage. Thus, we see that segment 2 represents 2.2% of all cases, but accounts for 4.5% of all respondents.

The next column displays the response rate for the associated segment (*score*). Thus, we see that segment 2 has the highest response rate (2.39%). The next highest response rate is 1.92% (segment 4).

The *score* represents the mean category score. By default, the category scores are '1' for the first category, and '0' for all others, so that the mean score corresponds to the % in the first category (responders in this example). To change the category scores,

- right click on the gains chart and select 'gains items' to bring up the gains chart control panel.



Note that a check-mark appears next to Responders to indicate that the default gains chart is presented.

- Click the Scores button, to bring up the gains chart category scores window.
- double click the score you wish to change, enter the replacement score and click the Replace button.
- Click OK after all the new scores have been entered.

To view the new gains chart based on the revised scores,

- click Responders in the Gains Chart control to remove the check-mark for the default gains chart.
- now click Responders once again in the Gains Chart control panel to restore the default gains chart.

The *index* column for a given segment measures the average response score for that segment relative to the average score for the total sample. The index score for segment 2 (208), which is computed as $(2.39\% / 1.15\%) \times 100$, means that the response rate for this segment was 108% higher than average.

Columns 8 through 13 in the gains chart present cumulative statistics. From the columns *labeled Cum: size, % of all, and score*, you can see that the three highest responding segments constitute 27.6% of the sample and have a combined response rate of 1.63%. The final column, *Cum: index*, measures the cumulative average response score for these segments relative to the average score for the total sample. For example, the index for the three best segments is 142 $(1.63\% / 1.15\%)$. Thus, the three best segments, taken together, responded at a rate 42% higher than average.

If you know the break-even response rate (or if the category scores reflect profitability), you can use gains charts to determine the segments to which you should mail future promotions. For example, suppose that when you take into account the cost of mailing and the gain from responders, you need a response rate of 1.45% to break even. Looking at the Gains chart above, (and assuming that this is your final segmentation), you would expect to make a profit if you mailed only the top two segments, since the score for the remaining households falls below the break-even level. Large savings could be gained by mailing only to segments with the highest response rates.

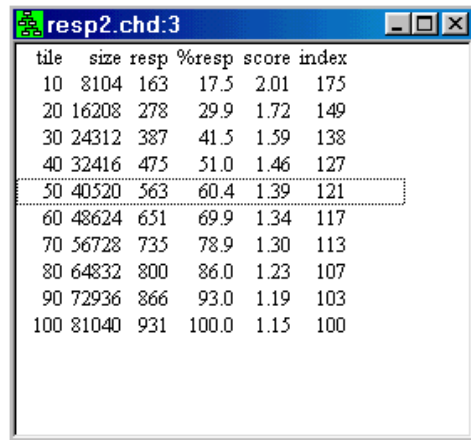
Summary Gains Chart

The summary gains chart summarizes the predicted response rate at various depths of the file. That is, the summary gains chart tells you the results that would be attained by targeting the best Q-percent of the file. This form of the gains chart is especially useful for comparing the results of 2 or more different CHAID trees. By default, the results are displayed in deciles.

To obtain a summary gains chart,

- click Summary on the (top) of the gains chart control panel.

The gains chart changes to the following:



tile	size	resp	%resp	score	index
10	8104	163	17.5	2.01	175
20	16208	278	29.9	1.72	149
30	24312	387	41.5	1.59	138
40	32416	475	51.0	1.46	127
50	40520	563	60.4	1.39	121
60	48624	651	69.9	1.34	117
70	56728	735	78.9	1.30	113
80	64832	800	86.0	1.23	107
90	72936	866	93.0	1.19	103
100	81040	931	100.0	1.15	100

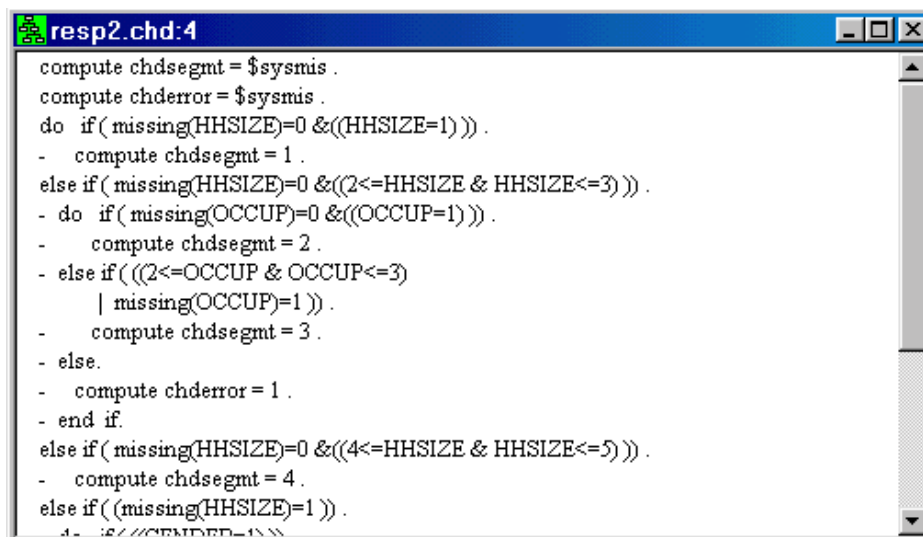
The *score* column shows that, the predicted response rate would be 2.01% if the best decile were mailed.

Scoring your file

You can obtain source code which will allow you to score your file with segment definitions.

- Select New Source from the Windows menu

A window appears containing SPSS if-then-else statements which compute the variable *chdsegmt* containing the CHAID segment number.



```
compute chdsegmt = $sysmis .
compute chderror = $sysmis .
do if ( missing(HHSIZE)=0 &((HHSIZE=1)) ) .
- compute chdsegmt = 1 .
else if ( missing(HHSIZE)=0 &((2<=HHSIZE & HHSIZE<=3)) ) .
- do if ( missing(OCCUP)=0 &((OCCUP=1)) ) .
- compute chdsegmt = 2 .
- else if ( ((2<=OCCUP & OCCUP<=3)
| missing(OCCUP)=1 ) ) .
- compute chdsegmt = 3 .
- else .
- compute chderror = 1 .
- end if .
else if ( missing(HHSIZE)=0 &((4<=HHSIZE & HHSIZE<=5)) ) .
- compute chdsegmt = 4 .
else if ( (missing(HHSIZE)=1) ) .
- compute chderror = 1 .
end if .
```

Crosstabulations

The *New Table* Window option displays a crosstabulation of the dependent variable (columns) by the current predictor variable (rows). You can control whether the table displays row percentages, column percentages, total percentages, or cell frequencies, and whether the table shows merged or unmerged categories of the predictor.

After-Merge Table

To view a crosstabulation showing row percentages for merged categories of HHSIZE at the top of the tree:

- Click the top (root) node of the tree diagram
- Select Window
- Click on New Table

Values in the Respondent column match the values displayed in each of the four HHSIZE nodes:

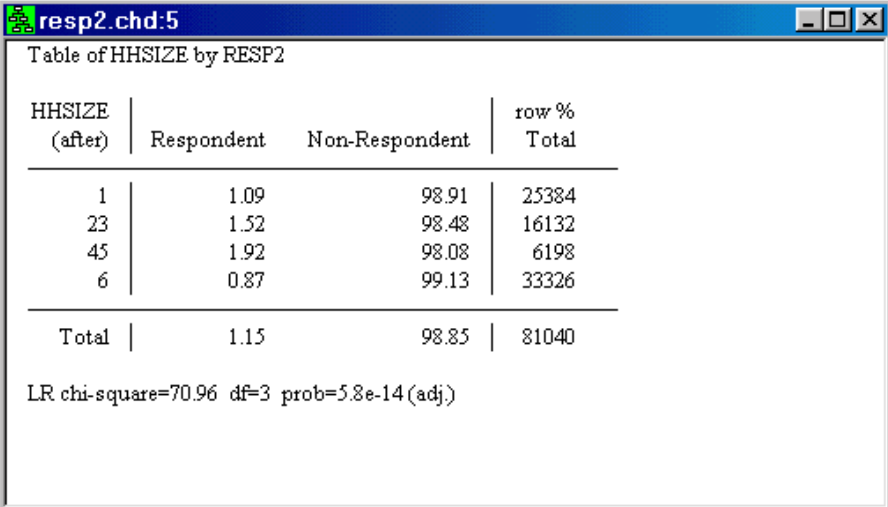
Notice that SI-CHAID merged categories 2 and 3, as well as categories 4 and 5.

The probability displayed in the bottom of the after-merge table, 2.7×10^{-15} , is adjusted for the fact that categories have been merged. The probability used by CHAID to rank predictors is the smaller of this adjusted probability and the probability associated with the table computed before category merging.

Before-Merge Table

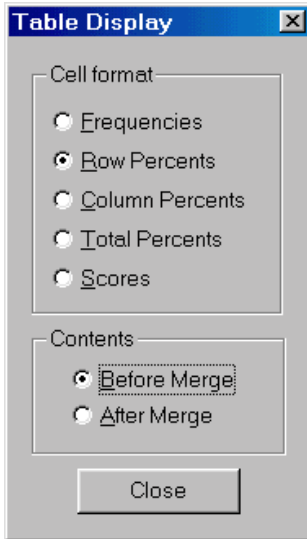
To view a row percentage table of HHSIZE by RESP2 for unmerged HHSIZE categories:

- Right-click on the Table and select *Table Items*.
- In the pop-up menu, click on *Before Merge*



HHSIZE (after)	Respondent	Non-Respondent	row % Total
1	1.09	98.91	25384
23	1.52	98.48	16132
45	1.92	98.08	6198
6	0.87	99.13	33326
Total	1.15	98.85	81040

LR chi-square=70.96 df=3 prob=5.8e-14(adj)



SI-CHAID automatically produces a table of row percentages before HHSIZE categories are merged, as shown below:

The table shows you the percentage of households in each HHSIZE category that responded to the promotion. For example, 1.09% of one-person households responded. Note that the total count in the lower right corner of the table (81,040) corresponds to the size of the highlighted node.

The table also displays the probability value (p value), a measure of statistical significance. The smaller the p value, the more statistically significant the predictor. The p value for HHSIZE before categories are merged is 4.4×10^{-14} (shorthand for 4.4×10^{-14} , a highly significant result). In fact, HHSIZE is the most significant of all the predictors. That is why the first split in the tree is based on household size categories.

Comparing Crosstabulations Before and After Merging

To see why some of the categories of HHSIZE have been merged, compare the Before- and After- Merge tables. SI-CHAID merged two-person and three-person households because their before-merge response rates (1.49% and 1.59%) are not significantly different. The combined response rate for the merged

HHSIZE (before)	Respondent	Non-Respondent	row % Total
1	1.09	98.91	25384
2	1.49	98.51	11240
3	1.59	98.41	4892
4	1.79	98.21	3187
Five or more	2.06	97.94	3011
.	0.87	99.13	33326
Total	1.15	98.85	81040

LR chi-square=71.79 df=5 prob=4.4e-14

categories is 1.52%. Similarly, SI-CHAID merges four- and five-person households, since the response rates for these subgroups (1.79% and 2.06%) are statistically indistinguishable. The combined response rate for the joint category is 1.92%.

Obtaining Frequency Counts

To obtain frequency counts before HHSIZE categories are merged

- Right-click on the Table and select *Table Items*.
- In the pop-up menu, click on *Frequencies*.

SI-CHAID automatically produces the table of frequency counts shown below;

HHSIZE (before)	Respondent	Non-Respondent	n Total
1	276	25108	25384
2	168	11072	11240
3	78	4814	4892
4	57	3130	3187
Five or more	62	2949	3011
.	290	33036	33326
Total	931	80109	81040

LR chi-square=71.79 df=5 prob=4.4e-14

The first row of the table indicated that 276 one-person households responded. The response rate displayed on the tree diagram (1.09%) is obtained by dividing the frequency by the total number of one-person households (25,384).

Growing the tree in Interactive mode

To explore your data in interactive mode, simply select any node of the tree you wish to analyze:

- Using the mouse or arrow keys, move to the HHSIZE = 23 node
- Right-click on the 23 node and select Select from the pop-up menu

The Select Predictors dialog box will come up. Three predictors show up as offering significant splits of this subgroup. They are ranked from most to least significant. At this point you may a) split the subgroup using the best predictor (OCCUP), b) select one of the other predictors to split on, or c) change the Detail level display selection to include variables that are not significant in the list of predictors.

- Highlight AGE and click OK to select it as the next predictor

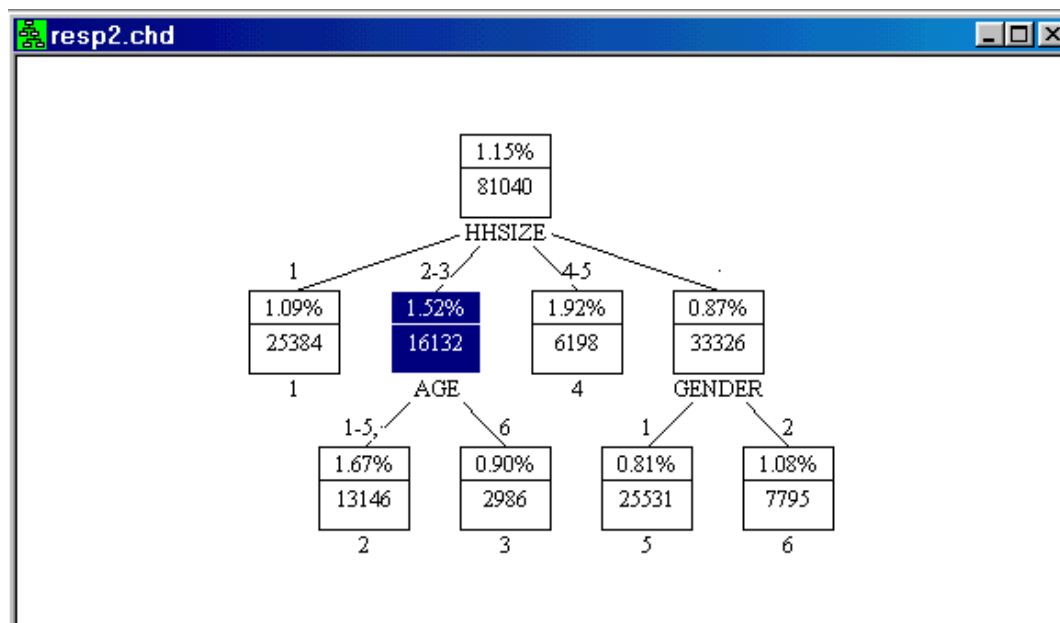
Id	Variable	p-Level	Categories	Groups
7	OCCUP	0.0085	4->2	1 2,3
1	AGE	0.012	7->2	1-5, 6
5	BANKCA...	0.012	2	1 2

Static

Detail Level
 Significant 2+ categories All

OK Cancel

The tree will now look as follows:



Rearranging Categories

- Right click and select Rearrange
- Select the 5 age ranges 18-64 as the 1st re-arranged category
- click the right arrow to move them to the right-most window

Id	Categories
6	65+
7	.

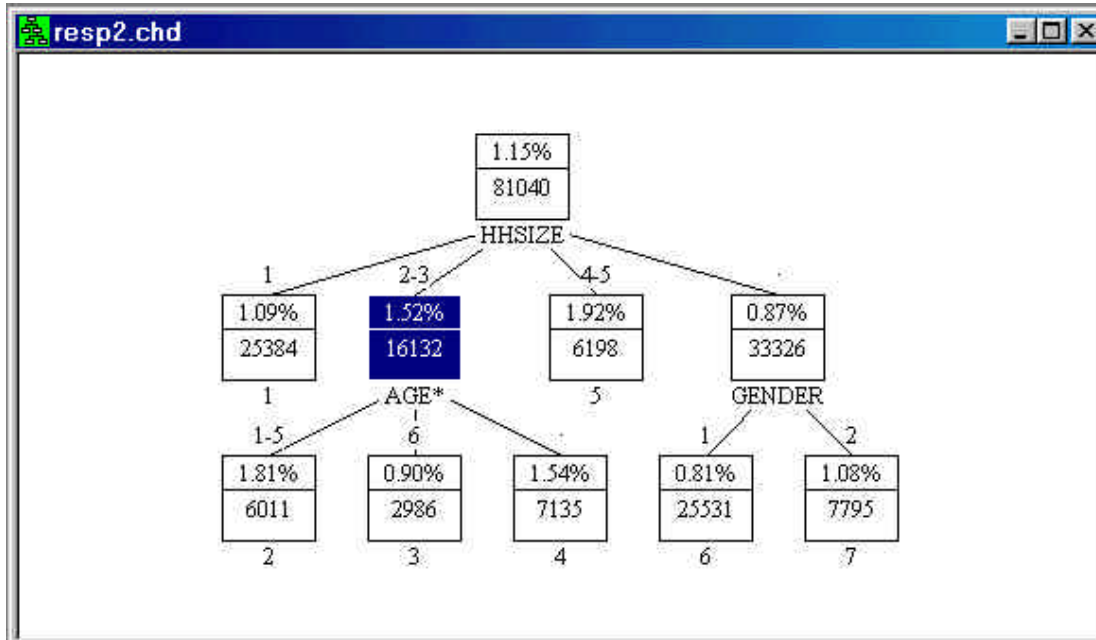
Next Prev

Id	Categories
1	18-24
2	25-34
3	35-44
4	45-54
5	55-64

OK Cancel Default Current Split All

- Click Next
- Select age 65+ as the 2nd re-arranged category
- click the right arrow
- click next
- Select the missing age group
- Click the right arrow
- Click OK

The rearranged tree will now look as follows:



SI-CHAID is designed as a useful tool to explore your data. There are no right or wrong trees. Feel free to explore your data as you wish.

