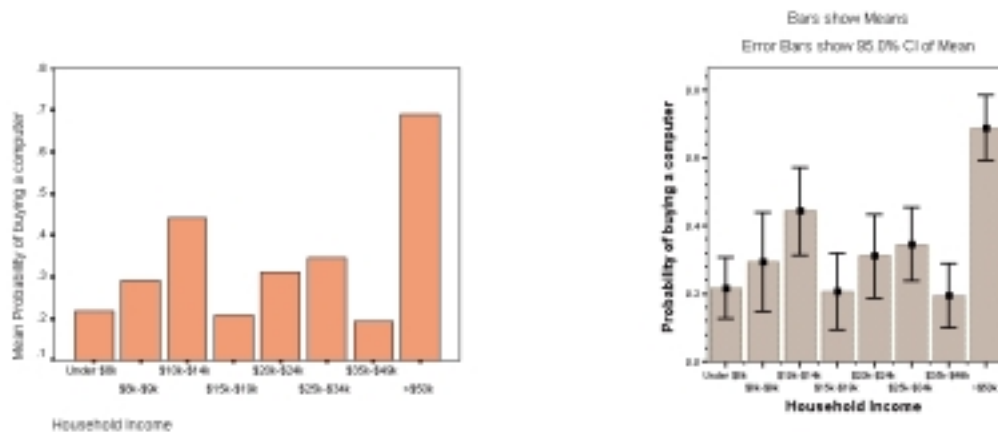


Determining the statistically significant segments in your database and using the information profitably.

R A Hoare, Ph.D.

Hoare Research Software, www.hrs.co.nz

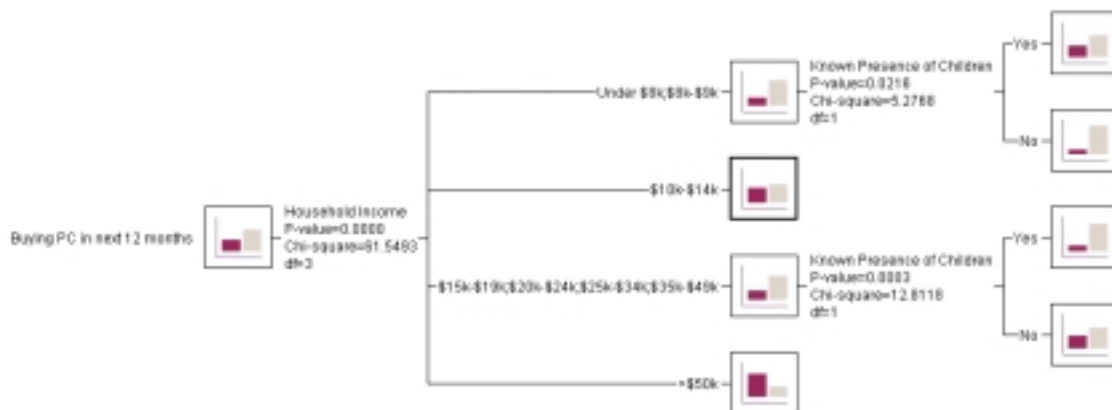
There are many ways to find out which segments of your database have historically been better targets for direct mail, bad credit risks, and so on. My talk covers some techniques that are based on statistical techniques, to try to convince you of the value of these.



Lets look at some data from a survey of people's intention to buy a computer.

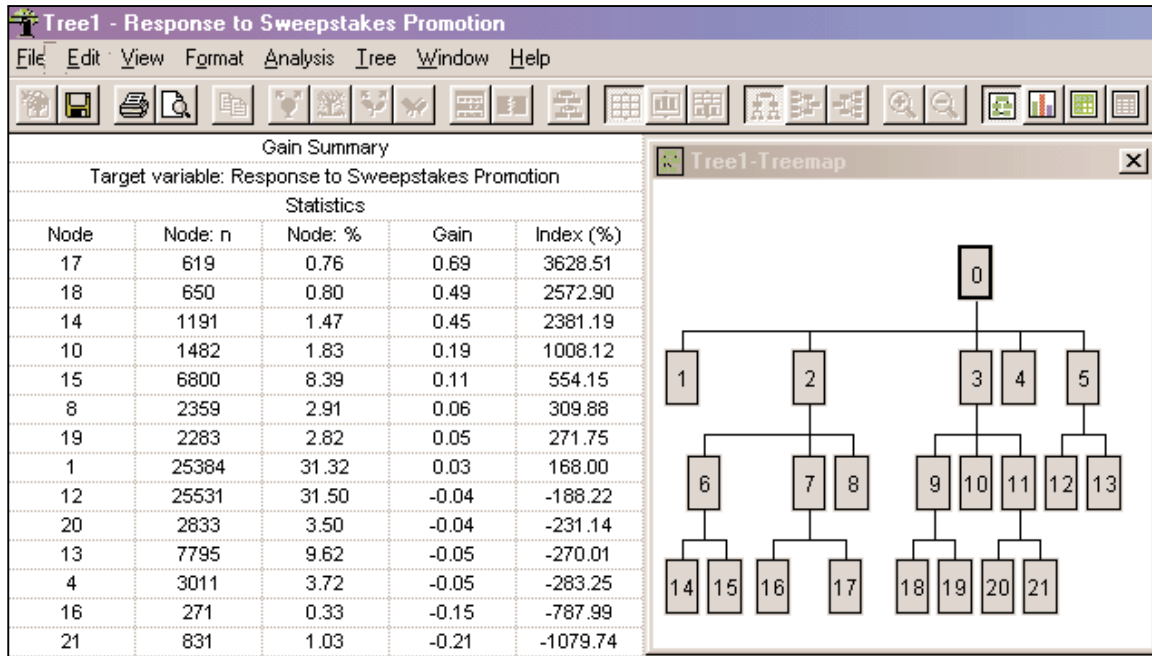
The left graph above makes it appear that some salary groups beside the highest one are likely buyers of computers. However, the graph on the right, with error bars, shows that next time you carry out a survey, or when you try to market to a new set of people, the pattern could be quite different.

In practice, you have data on a lot of aspects of your clients. Seeing combinations of these aspects, and taking into account statistical significance as well, is beyond most people. In the talk I demonstrate a technique called CHAID, that can systematically work through your data to find interactions, and present them in a "tree" diagram. Here is part of what you get from the above data and other information in the file.



The tree diagram (above) presents the proportion who are likely to buy in each group as a black bar. It shows that there are four statistically different groups in the income variable, and

that this variable is the best predictor of who will buy. Also, there is a dependence on the presence of children in the house. Low income groups with children, and middle income groups without children, are likely buyers.



Very often it is not enough just to know that one group is more likely to respond than another. Those who don't respond, or those who respond but don't buy, cost you money. If your activity is to be profitable, this cost has to be covered by the profit from those who do respond. In the table above, information on the response likelihood obtained from the tree diagram has been combined with the costs involved, to produce a table of gains. Node 17 has the greatest gain, and the figure means that on average you gain \$0.69 for everyone you contact in this group. Only 8 of the 14 end nodes are profitable. Note that nodes 1 and 12 have very large numbers of people in them. Although the profit per person is low in node 1, the total profit for that node is similar to the profit from other profitable nodes, because of the large number of people. The risk could be high, though, because the estimated profit per person is both small and uncertain.

The table to the right gives the proportions of the respondents for this data, for node 17. You can see that a very large proportion does not respond, in this dataset, but the profit from the rest makes it worthwhile.

Response to Sweepstakes Promotion : Node 17		
Cat.	%	n
Paid Respondent	2.42	15
Unpaid Respondent	0.16	1
Non-Respondent	97.42	603
Total	(0.76)	619

The selection rule, generated by the software, is

```

/* Node 17*/
IF (Number of Persons in Household NOT MISSING AND (Number of Persons in Household
> 1 AND Number of Persons in Household <= 2)) AND (Age of Household Head NOT
MISSING AND (Age of Household Head > 45-54 AND Age of Household Head <= 55-64))
AND (Household Income IS MISSING OR (Household Income > $10,000-$14,999))
THEN
  Node = 17
  Prediction = Non-Respondent
  Probability = 0.974

```

This rule (and others for the other profitable nodes) is used to decide how to select candidates from your own database or a purchased list in order to make your marketing effort worth while.

There are many more techniques that can be used to define profitable market segments and to estimate their value to you. If you contemplate using them I suggest you learn enough about them to satisfy yourself that the techniques are finding statistically significant groups, and that you can easily estimate the costs and benefits of using the segment information that you uncover.