Title: AN ITERATIVE APPROACH TO CLASSIFICATION ANALYSIS

Author: Cecily A. Gallagher
Howard M. Monroe
Joyce L. Fish

Biography: Ms. Gallagher is a Consulting Actuary with Tillinghast, a Towers Perrin company. She is a Fellow of the Casualty Actuarial Society and a Member of the American Academy of Actuaries. She holds a B.A. degree in mathematics from Trinity University, San Antonio, and a M.S. degree in statistics from Southern Methodist University.

Dr. Monroe is a Research Associate with USAA. Dr. Monroe holds a B.S. degree in mathematics and computer science from Texas A&I University and a Phd in statistics from Texas A&M University.

Ms. Fish is an Actuarial Research Analyst with USAA. Ms. Fish holds a B.A. degree in mathematics with a concentration in probability and statistics from Indiana University.

Abstract: This paper introduces a totally new approach to classification analysis. Part of its appeal stems from the fact that it provides a method for complying with Proposition 103's requirement that variables be considered in a specific order. Our paper is presented using private passenger automobile insurance; but, the technique could be used with any line of insurance.

The analysis is based on a statistical procedure known as CHAID, introduced in 1980. CHAID analyzes a single attribute of a population (the dependent variable) based on other attributes (predictor variables). CHAID iteratively subdivides the population into classes having significantly different values for the dependent variable.

This paper first describes some of the limitations of current classification analysis. We then introduce CHAID, and discuss how it overcomes these limitations. Because this is an introductory article, we have presented the statistical concepts of CHAID without fully explaining the underlying theory. Next, we present a CHAID analysis of "live" private passenger automobile experience. We then derive credibility weighted relativities for the CHAID classes. Finally, we briefly discuss actuarial and operational implications of a CHAID-derived class plan, and additional areas of needed research.

Historically, the objectives of classification analysis have been:

1. to subdivide the population into homogeneous groups whose loss costs can be predicted accurately; and

2. to measure each group's relative share of the total costs.

The more finely a population can be subdivided while still producing accurate individual group estimates, the better the allocation of costs.

Proposition 103 added a third classification objective: in determining a group's share of the costs, classification variables must be used in a specific order. To quote, "Rates and premiums ... shall be determined by application of the following factors in decreasing order of importance..."

Current actuarial techniques focus almost exclusively on the second objective; the measurement of the relative share of costs for predefined groupings of insureds. The analyst subdivides his population based on characteristics he believes are most significant and defines a relationship among these characteristics in terms of a premium determination equation. A single equation is used for the entire population and dictates, for example, whether merit rating is assumed to be independent of age, sex, marital status factors (implying an additive relationship) or whether a compounding effect (implying multiplication) is anticipated. The subdivisions of the data and form of the

premium determination equation are decided before examining the data. Data analysis then concentrates on the proper allocation of costs among the subdivisions in the context of this equation.

In allocating costs among the groups, the relative importance of a particular variable for the population as a whole is considered in one of two ways: either independent of all other variables, or in conjunction with certain variables and independent of the rest. For example, driver age, sex, and marital status are generally considered simultaneously, and may or may not be analyzed in conjunction with merit rating factors. When analyzed with merit rating, the potential for compounding (i.e. double counting) the effects of two factors working together is reduced. For example, higher class factors for youthful males respond to the fact that youthful males as a group are involved in more accidents. Merit rating factors (which include surcharges for prior accidents and violations) determined without consideration of the higher youthful factors may reflect some of the increased propensity for accidents recognized by the driver age factors. Considering age, sex, marital status, and merit rating simultaneously helps reduce such potential double counting.

However, all other distinctions (territory, vehicle differences, etc.) are either ignored in this analysis (by assuming that the groups are similarly represented in each class group and therefore differences "average out") or are reflected by using average relativities for the variables not under consideration.

Defining the data categories and form of the premium determination equation before the data is examined effectively reduces the analysis to determining the "best" set of factors derivable in a relatively narrow context, independent of the data. This imposes three serious limitations on the depth of the analysis.

First, by starting with a preconceived definition of the appropriate class groups, it is quite possible that the selected categories include groups that are not distinctly different in the population under review, and combine groups that are.

Second, because of the complexity of analyzing several factors simultaneously, more often than not, practicality limits acceptable premium determination equations to only the simplest forms. While variables which may be explaining the same effect with respect to accident involvement are considered concurrently, the interaction of these variables is generally ignored.

Third, and perhaps most important, by using a single equation, the procedures assume that all rating variables are equally important to the entire population. In other words, including age, sex, marital status, mileage, merit rating, territory and vehicle types in a single equation for all drivers implies that each of these factors carries the same significance for each type of driver. On the contrary, it is quite conceivable that variables may differ in their significance for different population segments. For example, the distinction between short and medium annual mileage may be meaningless for

drivers with a history of accidents. Or, completion of a driver training course may be irrelevant to young drivers with three or four moving violations.

Thus, although current actuarial techniques do address the historical objectives outlined in the previous section, they are limited in their effectiveness to identify significant class groupings and, hence, in distributing costs.

Finally, current procedures do not address the newest objective of classification analysis prescribed by Proposition 103. When variables are considered together, they are considered simultaneously, rather than in a specified order.

INTRODUCTION TO CHAID

CHAID is a statistical procedure that may help address these limitations. CHAID was first presented in an article entitled "An Exploratory Technique for Investigating Large Quantities of Categorical Data" by Dr. G.V. Kass in the 1980 *Applied Statistics*. The procedure, an offshoot of an earlier technique known as Automatic Interaction Detection (AID), uses the chi-square statistic as its primary tool.

CHAID is concerned with predicting a single variable, known as the dependent variable, based on a number of other variables, referred to as predictor variables. An analogous automobile insurance problem would be to predict either accident frequencies or pure premiums based on classification variables.

To quote from the *Applied Statistics* article, CHAID "... partitions the data into mutually exclusive, exhaustive, subsets that best describe the dependent variable". This is essentially the statistical equivalent of the first objective above for classification analysis.

CHAID is an iterative technique that examines the predictors (e.g. classification variables) individually and utilizes them in the order dictated by their statistical significance. The CHAID analysis is very straightforward and intuitive. CHAID first determines which of the prediction (classification) variables is most effective in distinguishing among different levels of the dependent variable within the population. It then partitions the population on the basis of significant categories (levels) of that variable. For example, if the analysis found that annual mileage is the best variable for distinguishing among loss frequencies, the population is divided using the significant mileage levels (e.g. under 2,500, 2,501-5,000, etc.). Each partition is then examined individually to determine which of the remaining variables is most effective in distinguishing among risks in that partition. The process is continued until all variables have been examined. Those that are significant trigger another division of the data; those that are not significant are discarded for that partition.

The result is similar to an inverted tree, with each branch identifying a significant subgroup of the population. Exhibit A illustrates an hypothetical result from a CHAID analysis.

CHAID directly responds to the limitations associated with current classification analysis described above. First, CHAID allows the data to define the appropriate class groups, thereby insuring that the groups identified pertain to the population under review. Second, because partitions are considered within the context of all previous factors, the interaction of all of the factors is automatically addressed correctly. Finally, because it is an iterative procedure, it provides an ordering of variables as required by Proposition 103. Although the ordering implied by the data may not be identical to that specified by Proposition 103, the technique may be altered to consider variables in a specific order.

However, CHAID does not address the second classification objective: the allocation of costs. CHAID merely identifies significant classes. Other actuarial techniques must be used to determine appropriate relativities based on these classes.

MATHEMATICAL DESCRIPTION OF CHAID

In this section, we present an overview of the mathematical concepts underlying CHAID. Dr. Kass' algorithm is available in a PL/1 program, and a general understanding of the procedure is sufficient to allow CHAID to be used effectively. For a more detailed explanation, we refer the reader to Dr. Kass' paper.

CHAID is basically an iterative four-step process:

1. Examine each predictor variable to determine which levels are significant in distinguishing among the differences in the dependent variable; compress all levels that are not significant.

2. Determine which of the predictors is the most significant in distinguishing among the dependent variable.

3. Subdivide the data by the levels of the most significant predictor. Each of these levels will now be examined individually.

4. For each level,

   a. examine the remaining variables to determine which levels are significant and compress all others.

   b. determine which predictor is the most significant and subdivide the data again by the levels of this variable.

5. Repeat step 4 for all subgroups until all statistically significant subdivisions have been identified.

CHAID involves a sophisticated application of the basic Chi-Square Contingency Test introduced in every basic statistics course. As a refresher, Exhibit B illustrates the mechanics of this test.

CHAID uses the chi-square statistic in two ways. First, it determines whether levels in the predictor can be merged together. Once all predictors are compressed to their smallest significant form, it then determines which predictor is the most significant in distinguishing among the dependent variable levels.

Variables

Before describing the process in more detail, we must explain the types of data that CHAID permits. The dependent variable must be divided into discrete categories. Predictor variables may assume three different forms: monotonic predictors (such as driver age) which have an implicit ordering, free predictors (such as territory) which have no implicit ordering, and floating predictors for which all but one of the levels follow a specific ordering. A floating predictor allows the use of a "missing" or "unknown" level (known as the floating level) in conjunction with an otherwise monotonic variable.

CHAID Algorithm

Assume the dependent variable has $d>1$ levels, and a specific predictor variable under analysis has $c>1$ levels. This data can be summarized in a $c \times d$ contingency table. The objective of step 1 of the CHAID analysis is to compress the rows of this $c \times d$ table to include only levels that are significantly different. In mathematical terms, we wish to reduce the $c \times d$ table to the most significant $j \times d$ table, $j=2,3,...c$. We then choose the $j \times d$ table that has the most significant chi-square statistic.

Permissible mergers of levels will depend on the type of predictor variable. Whereas levels of free predictors can be combined in any manner, levels of monotonic variables can only be merged with contiguous levels. Mergers of levels of floating predictors are also restricted to contiguous levels, with the exception of the floating category, which can either stand alone or be combined with any other groups.

Actual calculation of the best j x d table generally requires dynamic programming to examine all possible permutations. For monotonic predictors, the calculation is on the computational order of c^2. For free predictors, the solution is on the order of 2^c.

Clearly, such computations make dynamic programming an unrealistic approach for examining a large number of predictor variables. Thus, Dr. Kass has developed an alternative method, analogous to techniques used in stepwise and piecewise regression, which does not guarantee an optimal solution, but has produced very satisfactory results in practice. His algorithm is described in Appendix A.

Once the predictor levels have been compressed, the algorithm must determine the significance of the reduced contingency table. If there were no reduction to the table, the significance would be the complement of the probability of the computed value assuming (j-1)(d-1) degrees of freedom, which can be determined from any chi-square table.

However, if our contingency table has been reduced, the algorithm insures that the resulting table is the best possible for its size. The significance,

therefore, must reflect the fact that this table has not been considered in isolation, but in the context of all possible j x d tables. Therefore it is not sufficient to merely determine the significance associated with the computed chi-square statistic. Instead, Dr. Kass associates a significance related to the simultaneous consideration of all j x d tables. Using a probability theorem credited to Bonferroni, he computes a lower bound on the simultaneous significance of all j x d tables. This significance level is determined by multiplying the significance for the unreduced table by a factor known as the Bonferroni multiplier.

The multiplier corresponds to the number of ways that c levels can be reduced to j groups, and it differs for each type of predictor variable. Because only contiguous levels can be grouped for monotonic variables, the Bonferroni multiplier is $\binom{c-1}{r-1}$. The Bonferroni multiplier for free predictors, which can be grouped in any way, is: $\sum_{i=0}^{r-1} (-1)^i \frac{(r-i)^c}{i!\,(r-i)!}$

Not surprisingly, the multiplier for the floating predictors is the most complex, and takes the form: $\binom{c-2}{r-2} + r\binom{c-2}{r-1}$

A numerical example will help illustrate the use of the multiplier. A 4 x 5 unreduced contingency table with a test statistic value of 23.3 has a significance of .025 (assuming 12 degrees of freedom). However, if we start with a 6 x 5 contingency table and reduce it to dimensions 4 x 5, the same test statistic would reflect a significance of .10 (4 x .025) assuming a monotonic predictor variable. If we were using 5% significance level as our critical point, the unreduced table would be considered significant, but the reduced table would not.

The reader is referred to the original paper for a more complete explanation of the use of Bonferroni multipliers.

Thus, the final step in each iteration is to determine whether the predictor variable with the greatest significance is sufficiently significant to merit a partitioning of the data. If so, the data is subdivided based on the significant levels of that variable. If not, the process stops.

Appendix B presents a simple example of the analysis of one predictor variable based on the algorithm described in Appendix A.

The Data

We used CHAID to analyze a large private passenger data base containing records for all insureds written during a specific six month period. Each record contained the exposures earned (in months) and claims incurred during that time period. Exposures for each record vary from .5 to 6, since policies are written for a six month term. If policies were written evenly throughout the period, the average earned exposures would be 3.0. For our file, exposures averaged about 2.6, suggesting a little cyclicality. In the future this file will be extended to include writings over a full year.

The dependent variable was property damage claim counts ranging from 0 to 2. Insureds with more than 2 claims in a six month period were assumed to have only 2.

A variety of predictor variables were captured in the data base. For this analysis, we examined many of the traditional variables such as driver age, sex, marital status, annual mileage, accidents and convictions, as well as some less common variables such as number of vehicles on the policy, number of operators assigned to the vehicle, and original year insured with the company. Exhibit C describes the variables and the levels examined. Other variables, such as territory (or territory groups) and more refined subdivisions of some variables will be analyzed in the future.

For all but two variables, driver age and annual mileage, we used the full range of values available and allowed CHAID to designate the appropriate groupings. For age and mileage, we began with specific groups and allowed CHAID to further merge them if appropriate. Future research is planned to investigate each individual year of age, particularly for the younger drivers. We plan to use each driver's actual age, and allow CHAID to identify the appropriate groupings. Investigation of different mileage groupings is considered less important at this time.

We considered a number of overlapping variables relating to driver performance. We separately identified the total number of minor and major convictions and chargeable accidents assigned to the vehicle as well as the points assigned under the company's merit rating program. In our analysis, all driver performance variables pertain to the vehicle exposure. One area needing future research is the relative importance of the individual's record versus the policy or vehicle record.

Analysis

The CHAID program requires three input parameters: a significance level for partitioning a variable, a significance level for merging levels within a variable, and a minimum number of records for a cell to be considered for partitioning. We used a 5% significance level for partitioning; i.e. there is only a 5% probability that a partition determined by CHAID is spurious. The significance level for merging within a variable must be less than the level for partitioning. We used 4.9%. Finally, we required at least 500 records to be in a cell before it could be considered for partitioning.

We emphasize that although our analysis is based on actual experience, it should only be considered an illustration of the CHAID technique. The data is only a sample from a book of business and may not be representative of the entire population. The findings presented here are preliminary, and require considerably more analysis before accepted as definitive. Moreover, this is our first major application of CHAID with a live data set. Additional exploratory research is required to determine the consistency of the partitioning on similar samples, the consistency of the partitioning over time, and the sensitivity of the results to different input parameters.

Exhibit D presents the first three stages of partitioning for the entire population. It is interesting to note that before any partitioning, all variables with the exception of major convictions were significant at the 5% level when considered individually. Four variables in particular had a much greater significance level than the others: age of rated driver, annual mileage, marital status, and driver experience level.

Rated driver age was the most significant by a relatively substantial margin. Although we separated insureds into nine age groups, CHAID found only eight significant: the claim count distribution for drivers 60-64 was not significantly different from drivers 65-69.

After age, partitioning was triggered by different variables for different subgroups. For the youngest group, under age 21, the next most significant variable was merit rating points. For adults 30 to 49, the second most significant variable was the number of operators on the policy, which could be

considered a measure of exposure. For drivers between 50 and 59, annual mileage was the next most important factor, and for senior citizens, points again dominated all other factors.

Rather than discuss the entire population, we will concentrate on two groups: adults 30-49, the largest age subgroup, and drivers under 21.

Before focusing on these groups, three general aspects of the CHAID analysis bear comment. First, as can be seen from Exhibit D, although CHAID will not subdivide a cell with less than 500 records, it can produce a cell with very few records in it (e.g. operators 70-74, 6+ points). Because the chi-square test has a method for adjusting for varying sample sizes (i.e. the degrees of freedom) small cells are possible. Clearly, when we develop class relativities from this data, credibility considerations will eliminate such cells.

Second, at any iteration it is possible for two variables to be highly significant and their significance levels virtually identical. Although CHAID will select one of these for partitioning, it is conceivable that for another sample, CHAID may find the other variable slightly more significant. This is an aspect that bears additional investigation.

Finally, a variable may be used more than once in an analysis. For example, if CHAID divides a group based on no convictions versus at least one conviction, it is possible that after another partition of the latter group based on a variable other than convictions, CHAID may further subdivide on major convictions, e.g. separating 1-2 major convictions from more than 2.

Adults 30 - 49, 2 Operators on Policy

Exhibit E presents the complete analysis for adults 30-49 with two operators on the policy.

At the beginning of the process, the following factors were all significant at the 5% level:

Number of Vehicles on Policy
Annual Mileage
Vehicle Use
Number of Merit Rating Points
Sex
Marital Status
Number of Accidents
Number of Minor Convictions

However, when these factors were considered sequentially, CHAID reduced the number of significant variables to only three or four for each subgroup. This would suggest that there is considerable overlap in the discriminating power of the different variables; overlap that current techniques do not adequately address.

It can be argued that the CHAID analysis is largely influenced by our selection of the significance level, and selecting 10% or 15% instead of 5% would have produced additional partitions. This is supported to some extent by our data. If we had selected a 15% significance level, partitioning would have continued along three branches representing a fairly large percentage of this group; however, the statistics at this stage of the analysis suggest that it is unlikely that the branching would have continued too much farther.

Clearly one area that we will be examining in the future is the sensitivity of the analysis to this significance factor.

The importance of the significance criteria selected for merging levels is also evident. In this analysis, when levels were partitioned, they seldom resulted in more than two or three subdivisions. For example, in partitioning based on minor convictions for the 2,500 to 10,000 annual mileage group, CHAID only distinguished between no convictions and at least one. This lack of distinction among multiple convictions is most likely attributable, at least in part, to our merging significance level. Had we relaxed this to 10% or 15%, CHAID probably would have allowed more levels.

Given that this analysis will be followed by a credibility technique to establish relativities, it may be appropriate to use less restrictive parameters in CHAID to identify the possible existence of additional significant levels and allow the credibility procedure to determine whether they in fact are usable.

The paths and cells with a reasonable number of insureds identified by CHAID appear intuitively defensible to us. Some partitions were triggered by characteristics of this particular book. For example, widowed insureds comprise a unique market segment for this company and produced partitions that may not be generated from other populations. Other partitions were tied to the company's rating procedures. Clearly, the company's definition of merit rating points will affect the discriminating power of this variable, and the company's retention rate may affect the power of the "original year insured with company" variable.

We will leave the interpretation of the actual partitions to the reader, with the reminder that these analyses are preliminary and that we are not representing them as concrete findings.

Drivers Under 21

Exhibit F documents the full tree for drivers under 21. At the beginning of the analysis eight variables were significant at the 5% level:

Merit Rating Points
Vehicle Use
Good Student Discount
Inexperienced Operator Surcharge
Major Convictions
Number of Operators
Number of Vehicles
Minor Convictions

Surprisingly, sex and marital status were not significant, even at this stage of the analysis. It may be that the gap in claim frequencies between young males and females, which has been narrowing for some time, has finally closed, or that these variables would become significant if we were to further divide this group by individual age.

This analysis illustrates another interesting and intuitive aspect of CHAID. Although a variable is not significant at a specific step, it may become significant after partitioning. For example, even though the number of accidents is not a significant variable at the beginning of the process, it becomes important for drivers under 21 with 2-7 points.

The possibility of small cells is clearly evident in this group's analysis. CHAID produced cells of 4, 20, and 24 records as a result of partitioning a larger group. These will be of little or no use in a classification structure.

Once again, we leave interpretation of the actual partitions to the reader.

ALLOCATION OF COSTS

Although CHAID identifies significant classes, it does not offer a method for allocating costs among those classes. The allocation must be accomplished by other techniques.

Credibility Techniques

In the absence of credibility considerations, and still relying upon property damage frequency as our measure of relative risk, we would allocate costs based on the ratio of each CHAID class' average frequency to the average frequency of the total population, then adjust this ratio to a base class. Clearly, credibility cannot be ignored in light of some of the very small classes produced by CHAID.

At this stage, we have two issues to address. The first and more difficult is identifying a reasonable credibility complement. The second is determining the degree of credibility to assign to a cell that is less than fully credible. If CHAID produces consistent partitions for the population over time, eventually historical average frequencies could be used as complements.

However, because the data have not been analyzed in this fashion previously, there are no sources directly comparable at this time.

One complement we considered was the frequency estimate obtained from linearly regressing all of the cells on a particular level. In essence, each level would be "self-supporting". However, the CHAID process itself eliminates this as an alternative. CHAID effectively separates the observations on a particular level into cells clustering about a similar average. Thus, a linear regression of the data effectively fits the averages of the cells to a straight line, and reproduces the observed cell means.

In the absence of more representative complements, we used the average frequency of the cell directly above a particular cell's level as the complement. This is the average frequency that would have been assigned had there been no partitioning.

This criteria lends itself perfectly to the Bayesian credibility technique described by Phil Heckman in his paper "Credibility and Solvency" presented at the 1980 Discussion Paper Program. In his paper, Mr. Heckman outlined the calculations for estimating credibility under a nested (or heirarchical) structure in which a subgroup is assigned some credibility and its "parent" group assigned the remainder.

However, Mr. Heckman's paper is concerned with credibility associated with loss ratios rather than loss frequencies, and his credibility criteria is a function of premium rather than exposures. His credibility criteria takes the familiar form $Z = P/(P+K)$. In his application, P is the subgroup's premium, and K is estimated from the solution of a system of equations.

We believe that this "Heirarchical Bayesian" technique applies directly to our situation; however, adapting it represents a research project of its own. As an interim step, we offer the following less sophisticated approach.

Full Credibility

The first step in our procedure is to define a criteria for full credibility. Since we have limited our analysis to loss frequency, we have greatly simplified the problem, and can rely on some of the earliest efforts in credibility theory for a reasonable solution. We define full credibility as the number of observations required so that the probability is a least P that the true average frequency for the cell is within k% of the observed average frequency: i.e.

$$P[(1-k)\bar{X} < \mu < (1+k)\bar{X}] > P, \text{ where}$$

$\bar{X}$ represents the cell's observed average frequency,

μ represents the true average frequency, and

P and k are specified.

Under the assumption that each cell consists of independent and identically distributed random variables, with common mean μ and variance σ^2, then the Central Limit Theorem ensures us that the distribution of the variable $\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right)$ converges to a standard normal as the number of observations (n) increases.

Thus, we can rewrite our full credibility criteria as follows:

$$P\left[-\frac{k\bar{X}}{\sigma/\sqrt{n}} \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq \frac{k\bar{X}}{\sigma/\sqrt{n}}\right] \geq P$$

And, the number of observations required for full credibility will be:

$$n \geq \left(\frac{z_{p/2}\,\sigma}{k\bar{X}}\right)^2$$

Unfortunately, this criteria depends on an unknown quantity, σ^2. Historically, this problem was solved by simply assuming that the mean and variance of the population are equal. Then, the criteria for full credibility is totally determined by the selection of P and k. Selecting P = 90% and k = 5%, for example, leads to the popular 1,082 standard for full credibility.

However, the variance of the frequency distribution generally exceeds the mean. We believe that this fact should be reflected in the full credibility criteria. Thus, rather than assuming the mean and variance equal, we characterized the credibility criteria as a sampling problem from a normal population with an unknown variance, and addressed it using the t-distribution. Substituting the sample variance for the true variance, our full credibility criteria can now be written as: $n \geq \left(\frac{t_{p/2}\,S}{k\bar{X}}\right)^2$

Since, for large n, the t-distribution approaches the normal distribution, we again return to the normal distribution for our critical values. Thus the final form of our full credibility criteria is: $n \geq \left(\frac{z_{p/2}\, S}{k\, \bar{x}} \right)^2$

Exhibit G presents this derivation in greater detail.

This criteria differs from "classical" credibility criteria in two ways. First, because the variance is generally much greater than the mean for most of our cells, the number of observations required for full credibility is much larger than under the assumption of equal mean and variance. Second, there is no single credibility value: each cell's own variability determines its full credibility.

Partial Credibility

As mentioned above, the complement of credibility for a given cell is applied to the frequency of the cell immediately preceding the most recent partitioning. If the prior cell is not fully credible, the complement of credibility is assigned to that cell's credibility weighted frequency.

To determine partial credibility, we used the square root rule (i.e. the square root of the number of observations divided by the number required for full credibility).

An Application

Exhibit H summarizes the results of these procedures on one subgroup of the population: Adults 30-49, 2 Operators on Policy, driving 2,500 to 10,000 miles per year. Our full credibility criteria required the actual cell frequency to be within 10% of the observed average frequency with a probability of at least 90%. To determine class relativities, we would divide the credibility weighted frequencies shown in the last column of this exhibit by the population's average frequency and relate them to a base class.

This exhibit clearly demonstrates the impact of the different class experience on the full credibility criteria. All classes require substantially more observations than would be required under the criteria assuming equal mean and variance. Moreover, the number varies dramatically depending on the magnitude of the sample variance relative to the sample mean. For the class requiring the smallest number (1+ minor convictions, males, 1 vehicle) the mean and variance were .0509 and .19436, respectively. For the class requiring the largest number (0 minor convictions, widowed) the mean and variance were .0220 and .19033, respectively.

Assuming that the number of observations required for full credibility would be reasonably similar for other population subgroups, all other age groups with the possible exception of drivers over age 74, should be fully credible (Exhibit D). Thus, any credibility weighting would bring the less credible classes towards each age group, or a subgroup of that age group's frequency.

Minimum Bias Techniques

Although credibility techniques, and in particular the Hierarchical Bayesian credibility technique, seem to be the most obvious approach to allocating costs under CHAID, they may not be the only approach. An adaptation of minimum bias techniques commonly used in classification analysis today may provide an alternative allocation process. Whereas credibility techniques are focused on correcting for the variance in individual cells, minimum bias techniques attempt to minimize the bias of all individual classes while remaining unbiased in the aggregate. Minimum bias techniques require specification of a classification structure, a premium determination equation, and a minimization criteria (e.g., average absolute difference, squared differences). At this stage we have not explored the possibilities of using such a procedure with CHAID because of the difficulty of specifying a premium determination equation for our data. Depending on the classes derived from CHAID, however, specification of such an equation may be possible, making the minimum bias technique viable. Or, such procedures may be applicable to portions of a CHAID tree that are fully credible.

CONCLUSIONS

We believe that CHAID potentially could revolutionize classification analysis. CHAID is based on a more solid theoretical foundation than current classification ratemaking techniques, and solves many of the analytical problems that we have faced when trying to determine the combined effect of a variety of factors. Moreover, CHAID is much more intuitive than current techniques, which should eliminate much of the mystery surrounding classification analysis; mystery that is coming under increasing attack.

We have only opened the door to CHAID. We believe that our analysis demonstrates that CHAID can be used effectively with insurance data. Certainly much more analysis is needed before this technique can be used in practice. First, we must determine whether CHAID produces consistent class definitions over time, particularly when pure premiums rather than frequencies are used as the measure of relative risk. One additional problem to be resolved when we examine pure premiums is CHAID's requirement of a discrete dependent variable. Research will be required to determine appropriate groupings of pure premiums.

We must also find the most effective input parameters: the significance level for partitioning and merging, and we must develop effective techniques for allocating costs among the classes. Most important, we must demonstrate that the class relativities developed under CHAID do a better job of allocating costs than current techniques.

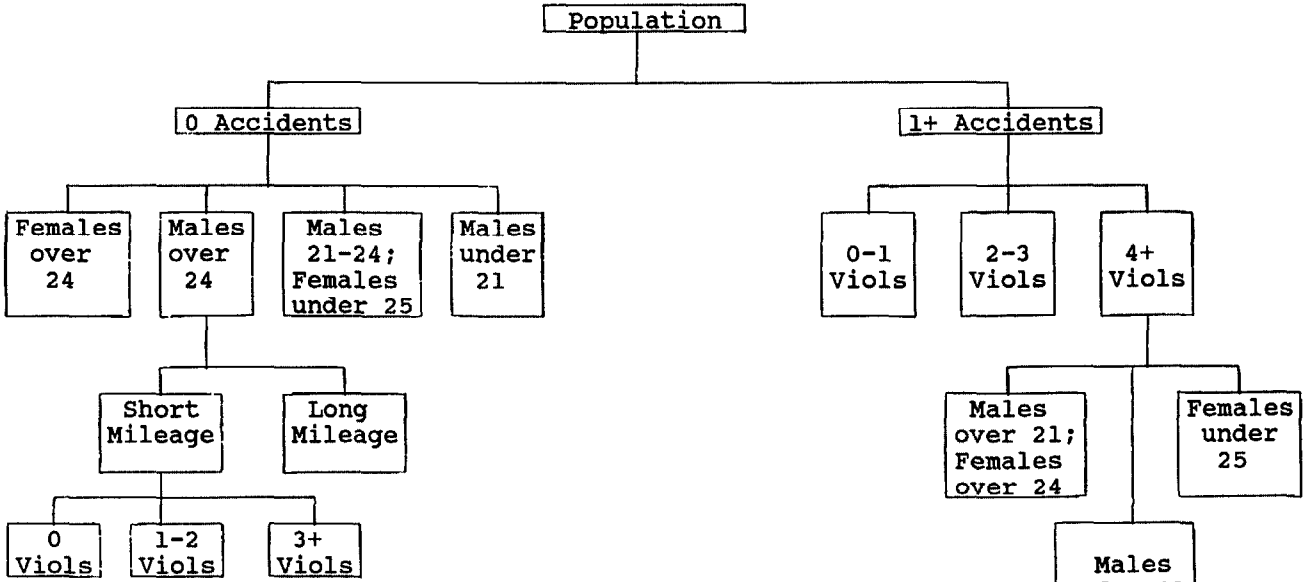
If CHAID is proven superior, it offers us a powerful tool for responding fully to questions regarding the relative importance of rating variables. For example, how does territory interact with other rating variables? What is the importance of the individual's driving record vis-a-vis the family's record? To what extent can mileage be used as a predictor of relative risk?

Converting to a CHAID-based classification structure could have far reaching consequences in many areas. New actuarial techniques for distributing statewide rate level indications may be needed, perhaps at the territorial level. CHAID analysis may alter underwriters' perspectives of particular

subgroups. Underwriting guidelines may have to be developed in conjunction with a new pricing approach to protect the company against adverse selection. Finally, creative ways of constructing rate manuals will be needed to make the classification structure understandable in the field.

We expect to be investigating CHAID for quite some time, and we hope that our article will interest others enough to join in the investigation.

CHAID ANALYSIS ILLUSTRATION
Schematic of Two Accident Subgroups



No other variables found significant

No other variables found Significant

Exhibit A

CHI-SQUARE CONTINGENCY TABLE
ILLUSTRATION

Hypothesis: Grades are independent of school attended

Observed Experience

School	Grades A	B	C	D	F	Total (R_i)
1	10	12	20	14	9	65
2	25	20	33	12	10	100
3	17	25	20	22	15	99
4	18	10	15	24	4	71
Totals (C_j)	70	67	88	72	38	335 (N)

Expected Experience ($R_i C_j / N$)

School	Grades A	B	C	D	F	Total
1	14	13	17	14	7	65
2	21	20	26	21	12	100
3	21	20	26	21	11	99
4	15	14	19	15	8	71
Totals	71	67	88	71	38	335

Test Statistic: $\sum_i \sum_j \frac{(obs_{ij} - exp_{ij})^2}{exp_{ij}} = 24.04$

Degrees of Freedom: 12 [(rows-1) x (cols-1)]

Significance Level: $\approx$ 0.021
(p-value)

CHAID DATA ELEMENTS

Data Type code is as follows: M - Monotonic
F - Free (nominal)
L - Floating (1, 2, 3)

FIELD NAME	DESCRIPTION VARIABLE PARTITIONS	TYPE
1. Incurred Claim Count	0 1 2 or more	M
2. Original Policy Year	1989 1988 1987 1985 or 1986 1984 or Prior	M
3. Number of Operators on Vehicle	1 2 3 4 5 6 7 8 9+	M
4. Number of Private Passenger Vehicle	1 2 3 4 5 6 7 8 9+	M
5. Annual Mileage	0-2,500 2,501-5,000 5,001-7,500 7,501-10,000 10,001-12,500 Over 12,500	M

FIELD NAME	DESCRIPTION VARIABLE PARTITIONS	TYPE
6. Vehicle Use	Adult: P - Pleasure Use DWS - Drive to Work Short DWL - Drive to Work Long B - Business Youth: P - Pleasure or Farm DW - Drive to Work	F
7. Number of Rating Points on Vehicle	0 1 2 3 4 5 6 7 8 or More	M
8. Sex	F - Female M - Male	F
9. Marital Status	M - Married S - Single D - Divorced W - Widowed L - Legally Separated	F
10. Inexperienced Operator	No Yes	F
11. Good Student Discount	No Yes	F
12. Driver Training Discount	No Yes	F
13. Defensive Driving Discount	No Yes	F
14. Age	Under 21 21 - 24 25 - 29 30 - 49 50 - 59 60 - 64 65 - 69 70 - 74 Over 74	M

FIELD NAME	DESCRIPTION VARIABLE PARTITIONS	TYPE
15. Number of Chargeable Accidents	0 1 2 3 4 5 6 7 8+	M
16. Number of Chargeable Major Convictions	0 1 2 3 4 5 6 7 8+	M
17. Number of Minor Convictions	0 1 2 3 4 5 6 7 8+	M

CHAID ANALYSIS
First Three Stages

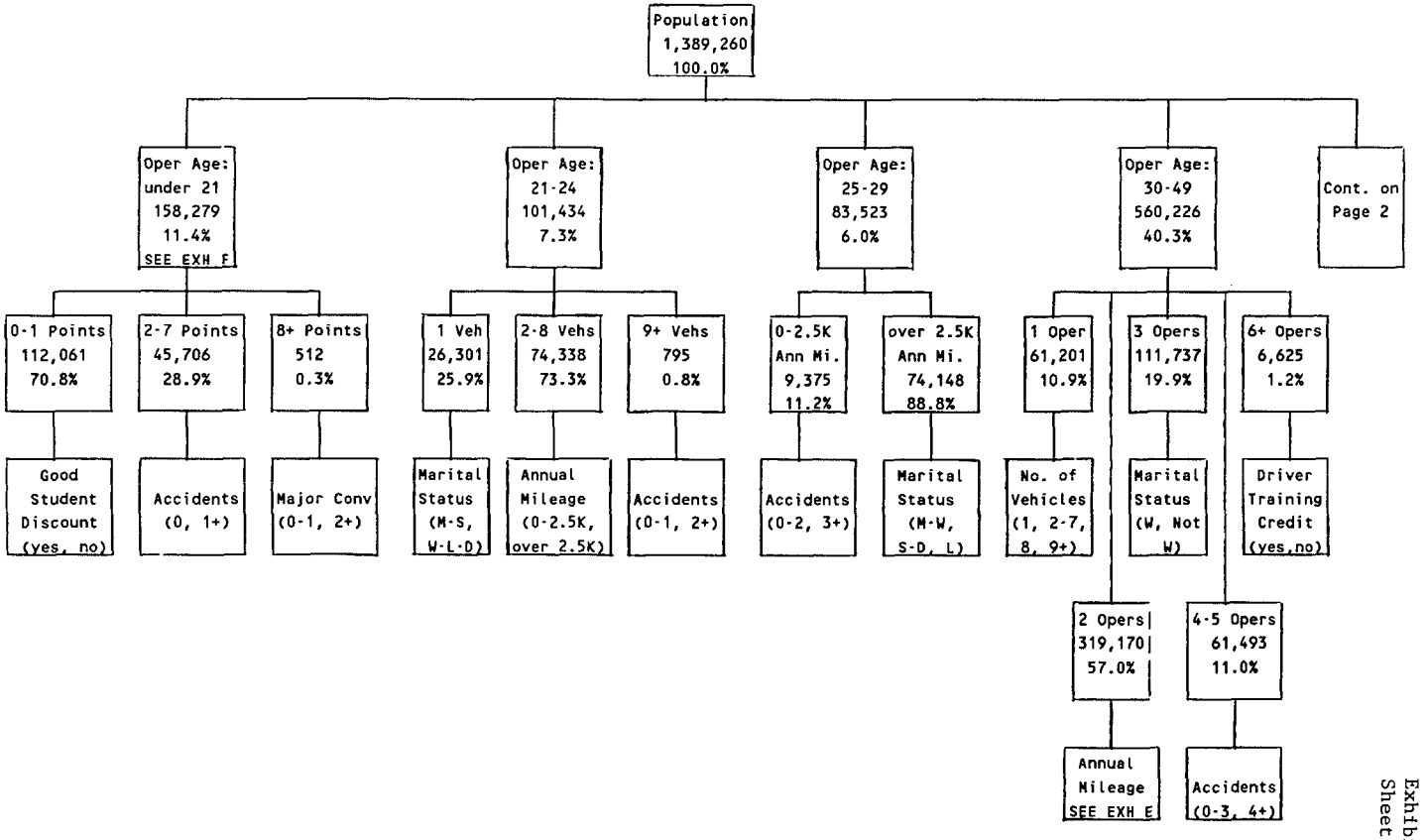


Exhibit D
Sheet 1

CHAID ANALYSIS
First Three Stages

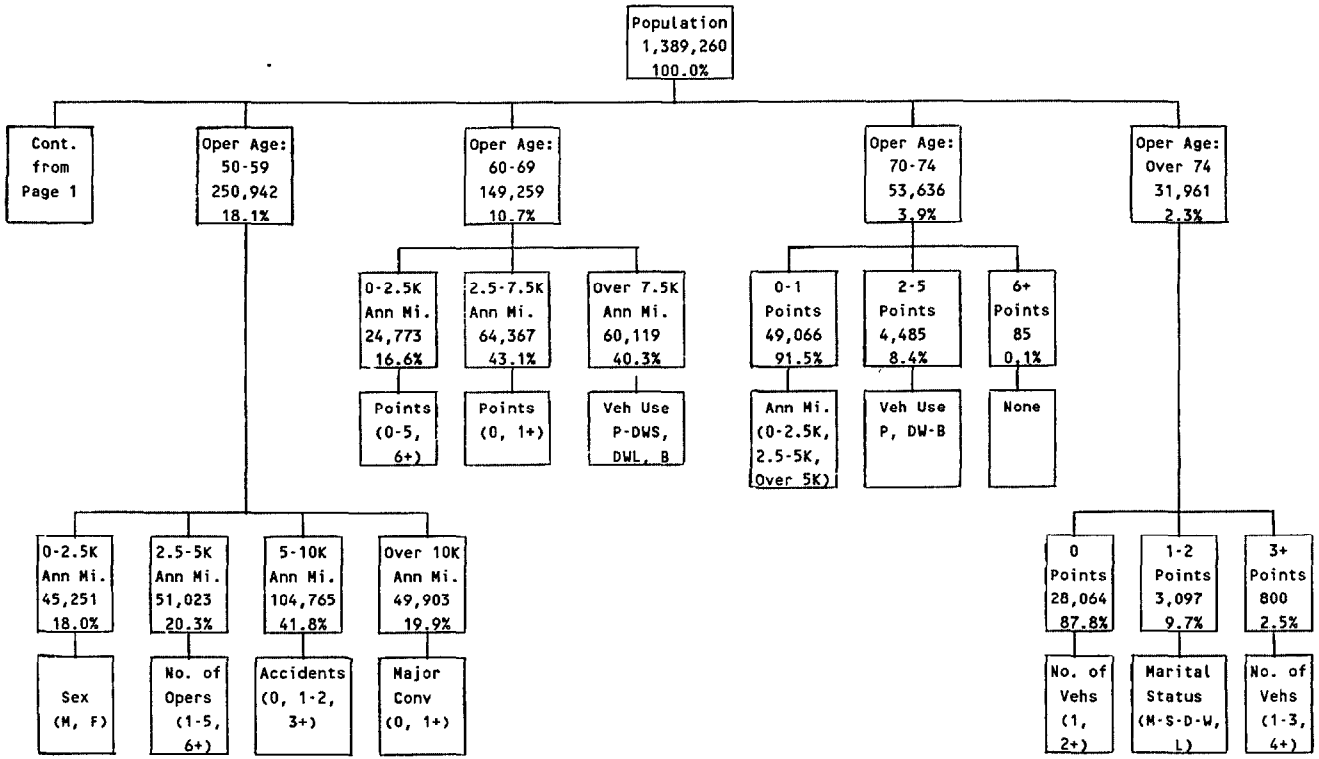
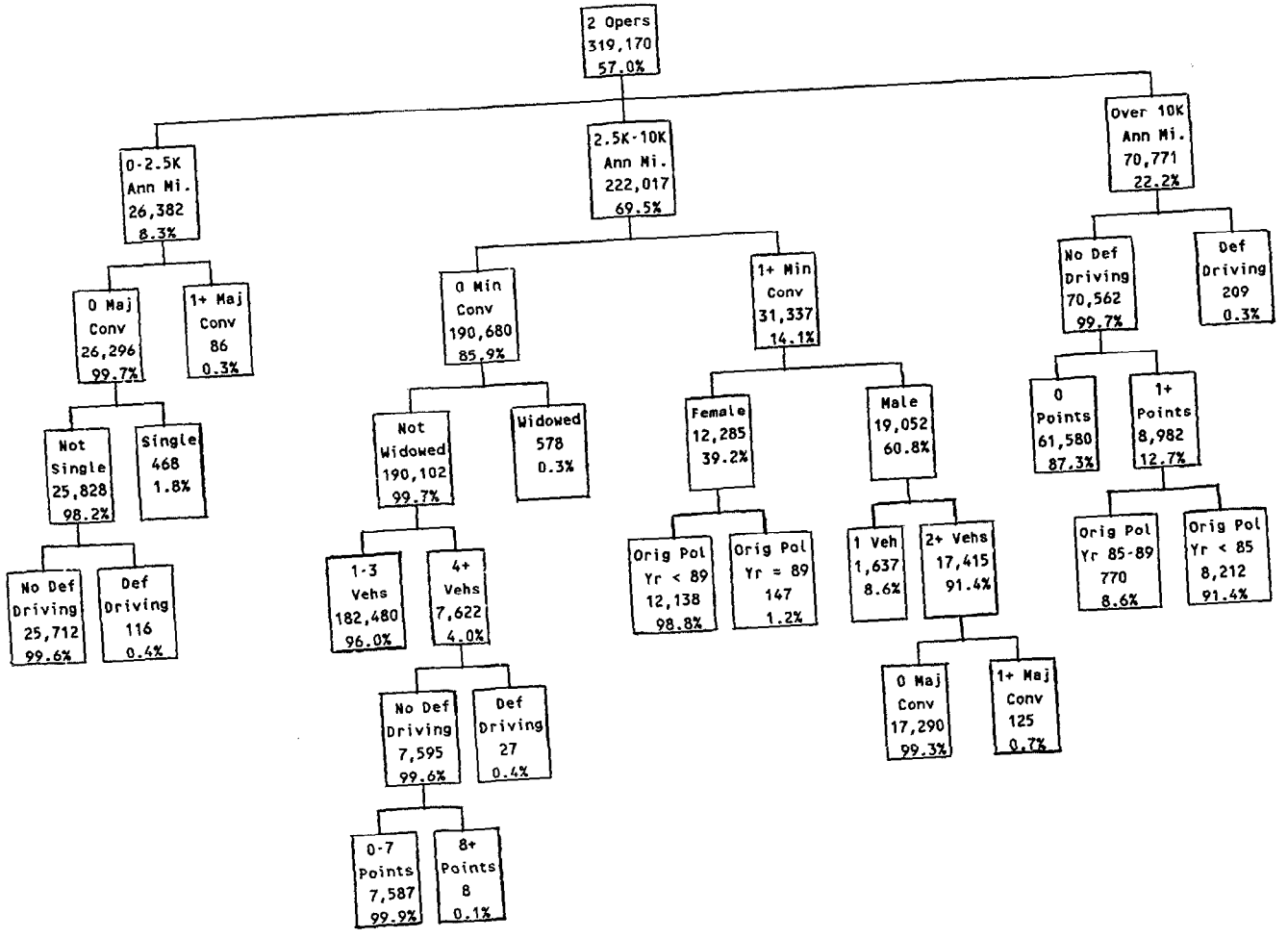
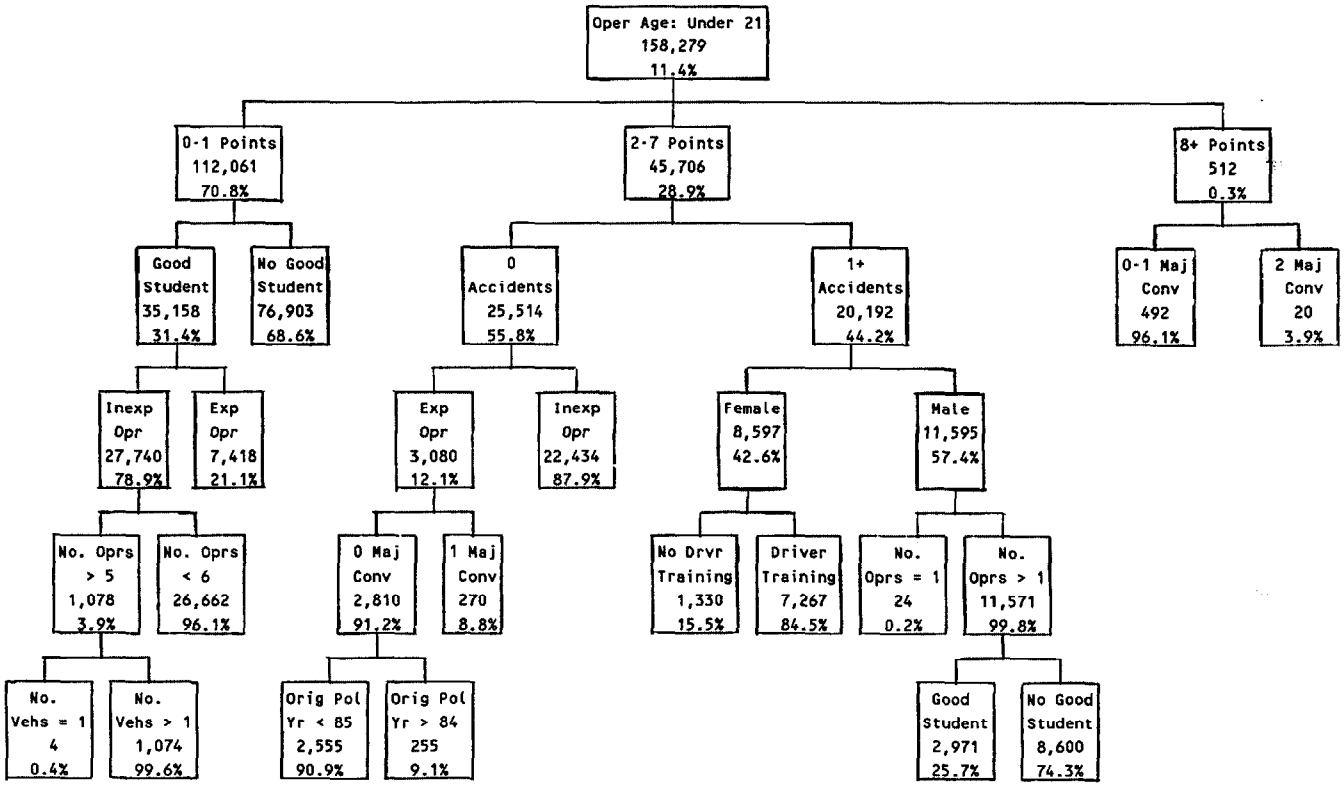


Exhibit D
Sheet 2

CHAID ANALYSIS
Drivers Age 30-49
Full Tree
(from Exhibit D)



CHAID ANALYSIS
Drivers Under 21
Full Tree
(from Exhibit D)



DERIVATION OF FULL CREDIBILITY CRITERIA

Criteria: $P[(1-k)\bar{X} < \mu < (1+k)\bar{X}] \geq P$, where $\bar{X}$ - sample mean
μ - population mean
P, k - specified

A. The criteria can be written as:

$$P[-k\bar{X} < \bar{X} - \mu < k\bar{X}] \geq P$$

$$= P\left[\frac{-k\bar{X}}{\sigma/\sqrt{n}} < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < \frac{k\bar{X}}{\sigma/\sqrt{n}}\right] > P, \quad \text{where} \quad \sigma^2 \text{- population variance}, \; n \text{ - number of observations}$$

According to the Central Limit Theorem,

(1) $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \quad \text{-->} N(0,1) \text{ as } n \text{ --> } \infty$

B. $\frac{(n-1)S^2}{\chi^2} \sim \chi^2_{n-1}$, where S^2 - sample variance
σ^2 - population variance
n - number of observations

We also know that

$$t_{n-1} = \frac{N(0,1)}{\sqrt{\frac{\chi^2_{n-1}}{n-1}}}$$

Thus,

(2) $$\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right) \Big/ \sqrt{\frac{(n-1)S^2}{\sigma^2(n-1)}} \sim t_{n-1}$$

DERIVATION OF FULL CREDIBILITY CRITERIA

C. So our full credibility criteria can be written:

$$P\left[\frac{-k\bar{x}}{\sigma/\sqrt{n}} \Big/ \sqrt{\frac{s^2}{\sigma^2}} \leq \left(\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}\right) \Big/ \sqrt{\frac{s^2}{\sigma^2}} \leq \frac{k\bar{x}}{\sigma/\sqrt{n}} \Big/ \sqrt{\frac{s^2}{\sigma^2}}\right] \geq P$$

$$= P\left[\frac{-k\bar{x}\sqrt{n}}{s} \leq t_{n-1} \leq \frac{k\bar{x}\sqrt{n}}{s}\right] \geq P$$

$$\Rightarrow \quad \frac{k\bar{x}\sqrt{n}}{s} \geq t_{p/2}$$

$$\Rightarrow \quad n \geq \left(\frac{t_{p/2}\, s}{k\bar{x}}\right)^2$$

Using the normal approximation to the t-distribution,

$$\Rightarrow \quad n \geq \left(\frac{z_{p/2}\, s}{k\bar{x}}\right)^2$$

Credibility-Weighted Frequencies

Adults 30-49, 2 Operators, 2,500-10,000 Miles

Class	Number of Records In Cell	Number of Records For Full Credibility	Frequencies In Cell	Frequencies Credibility Weighted	
2,500 - 10,000	222,017	60,986	0.0268	0.0268	
0 MINOR CONVICTIONS	190,680	63,484	0.0255	0.0255	
Not Widowed	190,102	63,628	0.0252	0.0252	
1-3 Vehicles	182,480	61,913	0.0254	0.0254	**
4+ Vehs	7,622	106,413	0.0220	0.0243	**
Widowed	578	33,322	0.0992	0.0352	**
1+ MINOR CONVICTIONS	31,337	49,382	0.0346	0.0330	
Females	12,285	46,426	0.0376	0.0354	**
Males	19,052	51,673	0.0326	0.0328	
1 Vehicle	1,637	20,300	0.0509	0.0379	**
2+ Vehicles	17,415	57,733	0.0309	0.0317	**

** Indicates end of branch

DESCRIPTION OF CHAID ALGORITHM

Below is the description of the CHAID algorithm as presented in Dr. Kass' 1980 article.

"Step 1. For each predictor in turn, cross-tabulate the categories of the predictor with the categories of the dependent variable and do steps 2 and 3.

Step 2. Find the pair of categories of the predictor (only considering allowable pairs as determined by the type of predictor) whose 2 x d sub-table is least significantly different. If this significance does not reach a critical value, merge the two categories, consider this merger as a single compound category, and repeat this step.

Step 3. For each compound category consisting of three or more of the original categories, find the most significant binary split (constrained by the type of the predictor) into which the merger may be resolved. If the significance is beyond a critical value, implement the split and return to step 2.

Step 4. Calculate the significance... of each optimally merged predictor, and isolate the most significant one. If this significance is greater than a criterion value, subdivide the data according to the (merged) categories of the chosen predictor.

DESCRIPTION OF CHAID ALGORITHM

Step 5. For each partition of the data that has not yet been analyzed, return to Step 1. This step may be modified by excluding from further analysis partitions with a small number of observations."

Source: "An Exploratory Technique for Investigating Large Quantities of Categorical Data", K.V. Kass, 1980, *Applied Statistics*

SAMPLE CHAID ANALYSIS OF ONE PREDICTOR VARIABLE

Dependent Variable: Claims Predictor Variable: Age of Driver (monotonic)

a. Before any merging

Age of Driver	Dependent Variable 0	1	2	3	Total	
Under 20	350	75	50	25	500	\ Least Significant
21-24	584	112	80	24	800	/ Test Statistic: 3.86
25-29	560	84	42	14	700	
30-49	3,440	340	140	80	4,000	
50-65	2,195	180	75	50	2,500	
Over 65	1,245	180	60	15	1,500	
Total	8,374	971	447	208	10,000	

b. After 1st merge

Age of Driver	0	1	2	3	Total	
Under 24	934	187	130	49	1,300	
25-29	560	84	42	14	700	
30-49	3,440	340	140	80	4,000	\ Least Significant
50-65	2,195	180	75	50	2,500	/ Test Statistic: 4.99
Over 65	1,245	180	60	15	1,500	
Total	8,374	971	447	208	10,000	

c. After 2nd merge

Age of Driver	0	1	2	3	Total
Under 24	934	187	130	49	1,300
25-29	560	84	42	14	700
30-65	5,635	520	215	130	6,500
Over 65	1,245	180	60	15	1,500
Total	8,374	971	447	208	10,000

All levels significantly different.

Significance of Variable: 0.00001 (Bonferroni adjusted)