

https://dspace.org/

Significance testing in, and some extensions of automatic interaction detection

Kass, Gordon Vivian

2013-06-12
http://hdl.handle.net/10539/12779

Downloaded from a DSpace repository. https://dspace.org/

SIGNIFICANCE TESTING IN, AND SOME EXTENSIONS OF AUTOMATIC INTERACTION DETECTION

Gordon Vivian Kass

A Thesis Submitted to the Faculty of Science
University of the Witwatersrand, Johannesburg
for the Degree of Doctor of Philosophy

Johannesburg, 1975

ABSTRACT.

This thesis is concerned with the technique of Automatic Interaction Detection (A.I.D.).

The first part of this thesis investigates the statistical significance of A.I.D. results. Primary concern is focussed on the explanatory power of a single predictor. The distribution of a test statistic which is asymptotically independent of the sample size is studied. An expression for its null asymptotic permutation distribution is derived, and critical values for both monotonic and free predictors in the special case of the uniform allocation of observations to its categories are tabulated. The evaluation of the asymptotic cumulative distribution function was performed by direct numerical integration for predictors with few categories, and by a specially modified Monte Carlo method in other cases. This Monte Carlo method is discussed in detail as it could be a potentially useful technique for integrating the multinormal distribution over closed regions. The asymptotic distribution of the test statistic is evaluated for some non-uniform allocations and compared with the tabulated results mentioned above. The exact null permutation distribution is derived in a special subcase of a three-category monotonic predictor and dichotomous dependent variable, and the results compared with the asymptotic distribution for various sample sizes. An upper bound to the critical values was obtained using Bonferroni's inequality, and found to form an excellent first approximation to the values of the special cases that were considered.

The second part concerns a number of extensions to A.I.D. which are discussed in the framework of a nominal dependent variable, thus extending conventional A.I.D. from the usual interval scaled dependent variable. The results of the first part of the thesis are used to produce a version of A.I.D. that

subdivides the data on the basis of the most significant predictor instead of the most explanatory predictor as commonly performed. The admission of significance testing during the A.I.D. analysis permits decisions on the number of subgroups to be formed from each parent group. Thus the traditional two-way splits are superceded by possible multi-way splits, or even no split at all if significance is lacking. Another innovation is the definition of a monotonic predictor with a single "floating" category, that is a single category, often representing missing information, that is not part of the monotonic scale. A generalized predictor including the free and monotonic types as well as the last mentioned type, is considered briefly. Repeated simulations of the extended technique provide further evidence for some of the claims made for the method; and yield timing estimates for the various computer routines. A numerical example concerning the analysis of matriculation examination boards, based on university student data, illustrates the scope of the extended A.I.D. A summary and an indication of possible future extensions and ideas for future research concludes the thesis.

For the Lord giveth wisdom:

(Proverbs 2:6)

ACKNOWLEDGEMENTS.

The help and encouragement of my two supervisors is gratefully acknowledged - to Mr. R.S.Hall head of the computer division at the National Institute for Personnel Research for some stimulating discussions and to Dr. D.M.Hawkins professor of statistics at the University of the Witwatersrand for his valuable advice.

My wife and daughter are lovingly thanked for their patient tolerance during the time-consuming preparation of this thesis.

CONTENTS.

PART I

CHAPTER 1. Description of Automatic Interaction Detection 1
1.1 History and general overview of the technique 1
1.2 Research on similar topics 6
1.3 Scope of this thesis - Part I 8
1.4 Notation 8

CHAPTER 2. The Test Statistic and its Distribution 11
2.1 Definition and general properties under H_0 11
2.2 General asymptotic distribution under H_0 13
2.3 Distribution in the monotonic case 16
2.4 Distribution in the free case 19
2.5 Note on distribution in cyclic case 23

CHAPTER 3. Computer Evaluation of Distribution 24
3.1 Evaluation by numerical integration 24
3.2 Evaluation by Monte Carlo methods 26
3.2.1 Partial analytic evaluation of multivariate normal integrals 27
3.2.2 Stratification outside the interior-hypercube .. 38
3.2.3 Generation of pseudo-random normal deviates .. 40
3.2.4 Results of Monte Carlo evaluation 43

CHAPTER 4. Tables and Discussion of Results 49
4.1 Cumulative distribution of K with equal allocation of observations to each category 49
4.2 Unequal allocation of observations to categories .. 51
4.3 Upper bound to the critical values 53
4.4 Evidence of applicability to non-infinite samples .. 57
4.5 Extension to several predictors 61

CHAPTER 5. Summary and Conclusion 65
5.1 Summary of results 65
5.2 Conclusions 66

PART II

CHAPTER 6. Introduction and Orientation 69
6.1 Introduction 69
6.2 Previous work in this area 72
6.3 Summary of contents of part II 77

CHAPTER 7. The Method of Analysis 79
7.1 Outline of the technique 79
7.1.1 Equivalence with A.I.D. in the case of a dichotomous dependent variable 86
7.1.2 Some properties of the technique 88
7.1.3 Convergence of the procedure 93
7.1.4 An alternative method (with an inherent weakness) 99
7.2 Types of predictor 101
7.2.1 Monotonic predictors 102
7.2.2 Free predictors 104
7.2.3 Floating predictors 110
7.2.4 Interleaved-monotonic predictors 114

CONTENTS (continued).

CHAPTER 8. Empirical Investigation 117
8.1 Simulation under the null hypothesis 117
8.2 Timing 119

CHAPTER 9. An Example 122

CHAPTER 10. Summary and Conclusions 129

LIST OF REFERENCES 133

APPENDIX A. Subroutine NORSP .. .. A 1

APPENDIX B. PL/1 Implementation .. .. B 1

APPENDIX C. Sample Output .. .. C 1

o0o

LIST OF TABLES.

3.1 Probability content of the largest standard interior-hypercube within ellipse $a_1X_1^2+a_2X_2^2=r$ 31

3.2 Probability content of the largest standard interior-hypercube within ellipsoid $a_1X_1^2+a_2X_2^2+a_3X_3^2=r$ 31

3.3 Probability content of standard interior-hypercube within a hypercube where all correlations are 0.1 .. 33

3.4 Probability content of standard interior-hypercube within a hypercube where all correlations are 0.3 .. 33

3.5 Probability content of standard interior-hypercube within a hypercube where all correlations are 0.8 .. 34

3.6 Comparison of the probability content of the best orientated hypercube found with the orientation obtained by Choleski's square root, in ten dimensions and all correlations equal to 0.3 37

3.7 Proportional stratification of region outside interior-hypercube 39

3.8 The regression coefficients α_c in the monotonic case for the model $\ln(1-D_c(k))=\alpha_c+\beta k$ 45

3.9 The regression coefficients α_c' in the free case for the model $\ln(1-D_c(k))=\alpha_c'+(\beta'+\gamma' c)k$ 46

4.1 Critical values in the monotonic case with equal proportions of observations within each category 49

4.2 Critical values in the free case with equal proportions of observations within each category 49

4.3 Asymptotic cumulative distribution D(k) in three monotonic categories 51

4.4 Asymptotic cumulative distribution D(k) in four monotonic categories 52

4.5 Asymptotic cumulative distribution D(k) in five monotonic categories 52

4.6 Bonferroni upper bound to the critical values for any allocation of the observations to categories in the monotonic case 55

4.7 Bonferroni upper bound to the critical values for any allocation of the observations to categories in the free case 56

7.1 Bonferroni multipliers in the free case 108

7.2 Bonferroni multipliers in the floating case 113

8.1 Number of significant results at 5% level in data simulating the null hypothesis 118

8.2 Time (in seconds) for each routine on an IBM S/370 Model 145 120

LIST OF FIGURES.

1.1 Categories of monotonic predictor 2
1.2 Cyclic predictor with 9 categories 3

2.1 Two dimensional region of integration for three free categories with equal numbers of observations within each category 21
2.2 Three dimensional region of integration for four free categories with equal numbers of observations within each category 22

3.1 Linear interpolation in the function $X=\mathfrak{F}^{-1}(U)$ 41
3.2 Linear interpolation before correction applied 42
3.3 Cumulative probabilities $D_C(k)$ in the monotonic case .. 45
3.4 Cumulative probabilities $D_C(k)$ in the free case 47

4.1 Cumulative null distribution of the test statistic K for three categories with equal numbers of observations within each category, for various sample sizes 62

7.1 Summary of data for one predictor 79
7.2 Flowchart depicting logic of analysis for a single predictor 83
7.3 Dichotomous dependent variable 86
7.4 A subtable of two rows 88
7.5 Four category example whose middle two categories have identical means 90
7.6 Two examples that demonstrate that the error of the approximation T=X+Y (see text) is not of constant sign 98
7.7 Example leading to a contradiction 100
7.8 Example of a monotonic predictor with seven categories 102
7.9 Example of splitting a monotonic predictor 103
7.10 Example of splitting a free predictor 105
7.11 Example of a free predictor whose categories cannot be ordered 109
7.12 Example of a compound group containing the floating category 111

9.1 Dendrogram of the analysis of matriculation boards .. 128

o0o

PART I

CHAPTER I. Description of Automatic Interaction Detection.

1.1 History and general overview of the technique.

Automatic Interaction Detection (A.I.D.) is the name of a technique to analyse dependencies in multivariate data. The present day procedure is described in Morgan and Sonquist (1963a, 1963b), and a similar idea which was not fully explored by the author is propounded in an earlier article by Belson (1957) in the field of biological classification. A.I.D. gained popularity with the availability of computer programs especially as described in Sonquist and Morgan (1964) - since the method is otherwise too tedious to employ unless only small quantities of data are analysed. However the main power of the technique comes from the analysis of large quantities of data; indeed the originators warn against its use on data sets of less than one thousand cases.

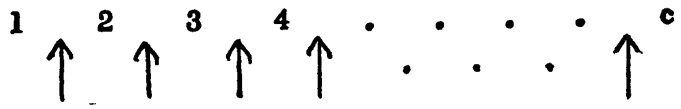
The aim of the method is similar to multiple regression in intent since it attempts to predict a single dependent variable on the basis of one or more predictors though without making the assumptions of linearity or additivity. The modus operandi, however, is completely different to regression analysis. In the words of the authors: Morgan and Sonquist (1963b) "It represents a simulation, of what careful researchers have done for years by hand or using an IBM sorter and tabulator when investigating a new set of data". While the dependent variable is . least ordinal as in stu.. regression, the predictors are considered

/2...

either ordinal or nominal: that is categorized variables which the A.I.D. literature terms monotonic or free. The procedure is to split the total set of data into two groups according to one or other predictor in order to minimize the within-group sum-of-squares, or equivalently to maximize the between-group sum-of-squares on the dependent variable. The predictor is used to define the split in that a subset of its categories itemizes one part of the split while the remaining categories determine the other part. Note that this implies that the two subgroups are disjoint.

If the predictor is monotonic then all the category numbers defining one subgroup of the split must be less than those defining the other subgroup. This is depicted in figure 1,1 which shows that if a monotonic predictor has c categories then it allows c-1 possible splits.

Figure 1.1 Categories of Monotonic Predictor



Position of Possible Split Points

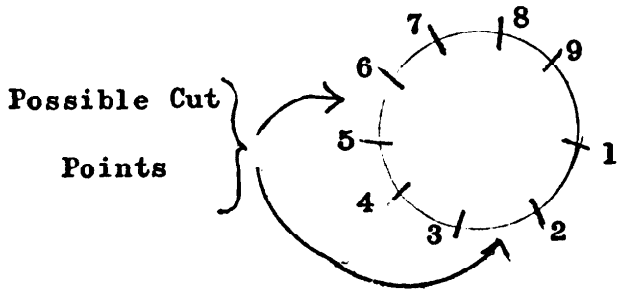
On the other hand a free predictor allows any grouping of its categories to define a subgroup of the total data. A c-category free predictor thus allows $2^{c-1} - 1$ possible splits since each category may designate one of the two parts of the split (2^c such possibilities); and it is immaterial which we call the first subgroup and which the second (hence divide 2^c by 2); and further the degenerate case of all

/3...

categories denoting the same subgroup must be excluded (hence subtract 1).

A cyclic predictor is also mentioned in the literature, but does not appear in use, so it is described here merely for the sake of completeness. This type of predictor is employed if its relationship with the dependent variable is considered J- or U-shaped, that is the categories at either extreme are considered more similar to each other than to those in the middle. The allowable clumping of categories is obtained by arranging the categories in order around a circle as in figure 1.2 and cutting the circle at two points so that the categories on each segment define a subgroup.

Figure 1.2 Cyclic Predictor with 9 Categories.



Note that if one of the cut points is between the first and last categories, then the predictor appears monotonic. The number of allowable splits obtainable with a c-category cyclic predictor is hence $\frac{1}{2}c(c-1)$, the number of ways of choosing 2 cut points out of the c possible positions on the circumference of the circle.

The A.I.D. algorithm finds the best (in terms of between-group sum-of-squares) possible method of splitting the data into two parts according to each predictor, which information is interesting in its own right, and then ascertains the

global best split across all the predictors. This best split is then actually carried out, and the analysis is repeated independently upon each of these two new groups. The process continues by further subdividing and analysing newly split groups until some terminating criterion is reached - usually in terms of some maximum allowable number of subgroups, or percentage variance explained, or the subgroups being too small to analyse further.

At this point we can amplify upon the name Automatic-Interaction-Detection. The "Automatic" signifies that the computer program makes its own selection as to the best split, and then acts on it in the hope that the user will agree. The "Interaction-Detection" indicates primarily that the technique does NOT assume an additive model, and that each subgroup is analysed independently, so that predictors that interact with a previously used predictor, will not have their effects smothered (if they should act in opposite directions within different subgroups) or unnecessarily included (if they should have an effect in one subgroup but not in another).

It may be noted that the version of the program under discussion would not in fact uncover an interaction unless one of the interacting predictors also had a strong main effect. A later version of the program, called AID-III, and described in Sonquist et al. (1971) has a look-ahead feature that could discover such interactions without main effects, by determining the reduction in sum-of-squares over two or

/5...

three stages of the analysis before actually performing the initial split. The authors claim (page 20) that they "have been unable to find a real-world example" where such an interaction is present; "And it is expensive in machine time to do all the extra calculation and looking".

A.I.D. has been used in a variety of fields, and in some cases unfortunately misused. These topics include Education (eg. Orr (1972)), Population Studies (eg. Ross and Bang (1966)), Market Research (eg. Assael (1970)) and Home Production (eg. Morgan et al. (1965)) to name but a few. An unfortunate use of A.I.D. on store performance and site selection given by Heald (1972) was criticized by Doyle (1973) who lists some of the other problems encountered in using A.I.D. and makes mention of other pitfalls into which an unwary researcher may fall if he is not given guidelines to differentiate between meaningful and idiosyncratic results.

Most users of A.I.D. are satisfied with the results in themselves and are happy to base the interpretation of their data on this alone. Only rarely is a mathematical model built on the basis of the A.I.D. except where a standard multiple linear regression is indicated, that is the predictors lie on an interval scale and their effects are additive. Since A.I.D. has not till now provided any test for the significance of its output, researchers basing their conclusions on this analysis alone have provided their own criteria as to whether they were noting a genuine effect or one merely due to chance - what Sonquist et al. (1971 p. 1) call the "never-never land of idiosyncratic results". This thesis is directed at partially alleviating

this defect, particularly to prevent abuse of the technique and to relieve the possible disquiet over the credence that may be placed in A.I.D. results.

1.2 Research on similar topics.

The primary work performed on evaluating the results of A.I.D. has been executed by Sonquist (1970). This empirical study of some artificially created data concludes that 0.6% of the total-sum-of-squares should be explained by each split. If this criterion is not reached, then it is recommended that such a group not be split. This particular criterion of 0.6% seems specific to his sample and choice of predictor variables rather than on an empirically derived estimate of some statistical percentage point. Cramer (1971) reviews this study and makes further criticisms of this appraisal of the technique. Heise (1971) also reviews this study and voices further cautionary remarks appertaining to the limitations of A.I.D.

No valid test of the significance of A.I.D. results appears to have been applied. The only criterion to test the difference between the two new groups formed by a split has been to fall back on the standard t-test. It is indisputable that insignificance of the t-test implies the split performed by the analysis is also not significant, however, significance of the t-test must be treated with extreme cau-

tion since A.I.D. by its very nature maximizes the value of the between-group t-statistic. In addition, there is usually more than one predictor variable, so that chance results are further capitalized upon.

A related problem is that of segmentation, where a set of variables ordered typically by time are to be clustered into homogeneous groups. For an example of the use of this technique see Hawkins and Merriam (1973). Here, each variable corresponds to the mean of the dependent variable within a category and the complete set of variables ordered on some secondary variable corresponds to a monotonic predictor. The general segmentation problem is restricted to finding the best binary division of the variables, and is extended from the usual homoscedastic variables to the heteroscedastic case. Work on this problem has been performed by Hawkins (1975). The major problems in segmentation are the location of the split point and the asymptotic results as the number of variables increases. Neither of these problems is of much concern in A.I.D. since predictors generally have less than ten categories (users are usually advised to combine categories if this is not so), and the major initial interest is whether or not a particular predictor is useful rather than in what way it should be employed.

MacNaughton-Smith (1963) has devised a similar technique to classify individuals. His work is a reduced version of A.I.D. in that he considers a dichotomous dependent variable together with a two category predictor. Our work will include his conclusions as a special case.

1.3 Scope of this thesis - Part I.

Part I of this thesis initiates an investigation into the statistical significance of results obtained by analysing data through A.I.D.

The next chapter proposes a test statistic that is asymptotically independent of the sample size. Its properties and asymptotic distribution under the null hypothesis are derived for any one stage in the analysis. Chapter 3 concerns the computer evaluation of the distribution of the test statistic and introduces some special techniques to evaluate multivariate normal integrals. Chapter 4 provides tables of critical values in a special case and compares various other situations with the special case. The chapter concludes with a discussion of the possible extensions of the results to various problems that are not considered elsewhere in this thesis. The last chapter in this part summarizes the results and considers their practical application. In the conclusion further possibilities are contemplated and part II of this thesis is introduced.

1.4 Notation.

The initial problem relates to splitting an arbitrary group (either the original total group or a subset of this as produced by the analysis) into two subgroups.

/9...

Let this group under consideration have N observations, with mean and variance of the dependent variable $\bar{X}$ and S^2 respectively. Let a particular predictor variable have c non-empty categories which can be numbered 1, 2, 3, ... ,c, and let N_i, $\bar{X}_i$, and $f_i = N_i/N$ be the number, mean and proportion of observations within category i respectively.

For an arbitrary permissible split i, let M_{1i} and $\bar{Y}_{1i}$ be the number and mean of the observations in the one subgroup, while M_{2i} and $\bar{Y}_{2i}$ denote the number and mean of the observations in the other subgroup.

The proportion of the sum-of-squares explained by this split is given by

$$P_i = \frac{\text{Between subgroup sum of squares}}{\text{Group sum of squares}}$$

$$= \left(M_{1i}(\bar{Y}_{1i} - \bar{X})^2 + M_{2i}(\bar{Y}_{2i} - \bar{X})^2 \right) / NS^2 \qquad \ldots\ldots\ldots\ldots (1.1)$$

and since $\bar{X} = (M_{1i}\bar{Y}_{1i} + M_{2i}\bar{Y}_{2i})/N$

and $M_{1i} + M_{2i} = N$

$$P_i = \left(M_{1i}(\bar{Y}_{1i}(N-M_{1i}) - M_{2i}\bar{Y}_{2i})^2 / N^2 \right.$$
$$\left. + M_{2i}(\bar{Y}_{2i}(N-M_{2i}) - M_{1i}\bar{Y}_{1i})^2 / N^2 \right) / NS^2$$

$$= \left(M_{1i}M_{2i}^2/N^2 + M_{2i}M_{1i}^2/N^2 \right) (\bar{Y}_{1i} - \bar{Y}_{2i})^2 / NS^2$$

$$= M_{1i}M_{2i}(\bar{Y}_{1i} - \bar{Y}_{2i})^2 / N^2S^2 \qquad \ldots\ldots\ldots\ldots (1.2)$$

/10...

and since $\bar{Y}_{2i} = (N\bar{X} - M_{1i}\bar{Y}_{1i}) / M_{2i}$ it follows from 1.1 that

$$P_i = \left(M_{1i}(\bar{Y}_{1i} - \bar{X})^2 + (\bar{X}(N-M_{2i}) - M_{1i}\bar{Y}_{1i})^2/M_{2i} \right) / NS^2$$

$$= \left((M_{1i}M_{2i} + M_{1i}^2)(\bar{Y}_{1i} - \bar{X})^2/M_{2i} \right) / NS^2$$

$$= M_{1i}(\bar{Y}_{1i} - \bar{X})^2 / M_{2i}S^2 \qquad \ldots\ldots\ldots\ldots\ldots (1.3)$$

The test statistic, whose null distribution is studied in the next chapter, is based on the set of P_i.

/11...

CHAPTER 2. The Test Statistic and its Distribution.

2.1 Definition and general properties under H_0.

The test statistic that will be considered is

K = Group size x maximum proportion of sum-of-squares explained over all permissible methods of splitting the group into two.

$= \max_i \{K_i\}$

$= \max_i \{NP_i\}$ in terms of our earlier notation.

Note that $K_i = NP_i$ is the usual ANOVA criterion for comparing two groups. The range of i depends on the type of predictor and summarizing the remarks made in chapter 1:

Monotonic Predictor - $i = 1$ to $c-1$

Free Predictor - $i = 1$ to $2^{c-1} - 1$

Cyclic Predictor - $i = 1$ to $\frac{1}{2}c(c-1)$

Defining M_1, M_2, $\bar{Y}_1$ and $\bar{Y}_2$ as the characteristics corresponding to M_{1i}, M_{2i}, $\bar{Y}_{1i}$ and $\bar{Y}_{2i}$ for that split that maximizes P_i, it follows from 1.2 and 1.3 that

$$K = M_1 M_2 (\bar{Y}_1 - \bar{Y}_2)^2 / NS^2 \quad \ldots\ldots\ldots\ldots\ldots\ldots (2.1)$$

$$= M_1 N (\bar{Y}_1 - \bar{X})^2 / M_2 S^2 \quad \ldots\ldots\ldots\ldots\ldots\ldots (2.2)$$

It is immediately obvious from 2.1 that K is non-negative and is zero if and only if the two new subgroups have identical means $\bar{Y}_1$ and $\bar{Y}_2$. The larger the difference between the two groups, the larger will be the test statistic

/12...

and hence the critical region for the rejection of the null hypothesis that the two subgroups are identical is the right-hand tail of the distribution of K.

When there are only two categories, there is but a single permissible method of splitting the group and all the types of predictors: monotonic, free and cyclic, become identical while the maximizing operation falls away. If the observations are normally distributed, the test statistic K has the familiar variance-ratio distribution as is obtained when testing the difference between two groups by an Analysis of Variance. Equivalently, K is the square of the usual t-statistic derived when testing for a significant difference between the two new subgroups. As the sample size increases the variance-ratio distribution tends to a chi-square distribution (or equivalently, the t-distribution tends to normality whose square is again chi-square) and hence in this case of a two category predictor, the test statistic K tends to chi-square with one degree of freedom. This is a generalization of the result discussed by MacNaughton-Smith (1963) in the special case of a dichotomous dependent variable.

Each K_i obtained by a specified method of splitting the group is distributed as above - and the test statistic K is a maximum of these K_i. Unfortunately these K_i are not statistically independent and hence the distribution is the maximum of a set of correlated chi-square variates.

/13...

2.2 General asymptotic distribution under H_0:

In order to provide a flexible approach the permutation distribution of K will be considered, and this will allow later extension to a distribution-free analogue of A.I.D.

Under the null hypothesis that the predictor is completely unrelated to the dependent variable, each allocation of the N observations to the different categories given the number N_i within each category is equiprobable. With this assumption the probability distribution of the mean $\bar{X}_i$ in category i with N_i observations is asymptotically normal, and by elementary finite sampling theory (see for example Cochran (1963)) has mean $\bar{\bar{X}}$ and variance $(1-f_i)S^2/N_i$ (where f_i, the proportion of observations within the i^{th} category, is the sampling fraction). The speed of convergence of the distribution of $\bar{X}_i$ to normality depends on the distribution of the raw scores: Bradley (1968) considers examples where the rate of convergence is much slower than expected.

The joint distribution of the $\bar{X}_i$ (i = 1, 2, ..., c) is then asymptotically a degenerate multinormal distribution in c-dimensions with mean vector

$$\underset{\sim}{X} = (\bar{\bar{X}}, \bar{\bar{X}}, \ldots, \bar{\bar{X}})'$$

and covariance matrix $\sum = (\sigma_{ij})$

where $\sigma_{ii} = (1-f_i)S^2/N_i$ as above . . . . . (2.3)

/14...

$$\text{and} \quad \sigma_{ij} = \text{Cov}(\bar{X}_i, \bar{X}_j) = E(\bar{X}_i \bar{X}_j) - \bar{\bar{X}}^2$$

$$= E\left(\bar{X}_i E(\bar{X}_j \mid \bar{X}_i) \right) - \bar{\bar{X}}^2$$

$$= E\left(\bar{X}_i (N\bar{\bar{X}} - N_i \bar{X}_i)/(N - N_i) \right) - \bar{\bar{X}}^2$$

$$= N\bar{\bar{X}} E(\bar{X}_i)/(N-N_i) - N_i E(\bar{X}_i^2)/(N-N_i) - \bar{\bar{X}}^2$$

$$= N\bar{\bar{X}}^2/(N-N_i) - N_i\left((N-N_i)S^2/NN_i + \bar{\bar{X}}^2 \right)/(N-N_i) - \bar{\bar{X}}^2$$

$$= \bar{\bar{X}}^2\left(N/(N-N_i) - N_i/(N-N_i) - 1 \right) - S^2/N$$

$$= 0 - S^2/N = -S^2/N \qquad \ldots\ldots\ldots\ldots (2.4)$$

To obtain the distribution of K, let

$$G_i = \left(f_i/(1-f_i) \right)^{\frac{1}{2}}$$

and make the linear transformation

$$V_i = N^{\frac{1}{2}} G_i (\bar{X}_i - \bar{\bar{X}})/S \qquad i = 1, 2, \ldots, (c-1)$$

so that asymptotically the joint probability distribution of the V_i given the proportion of observations f_i within each category is

$$h(\{v_i\} \mid \{f_i\}) \sim N(0, \Sigma_v)$$

where the covariance matrix Σ_v has unit variances

$$\sigma_v(i,i) = NG_i^2 (\text{Var}(\bar{X}_i))/S^2 = N\left\{ f_i/(1-f_i) \right\}\left\{ (1-f_i)S^2/N_i \right\}/S^2$$

$$= 1 \quad \text{from 2.3} \qquad \ldots\ldots\ldots\ldots (2.5)$$

$$\text{and} \quad \sigma_v(i,j) = (N^{\frac{1}{2}} G_i/S)\ \text{Cov}(\bar{X}_i, \bar{X}_j)\ (N^{\frac{1}{2}} G_j/S)$$

$$= NG_i G_j(-S^2/N)/S^2 \qquad \text{from 2.4}$$

$$= -G_i G_j \qquad i \neq j \qquad \ldots\ldots\ldots\ldots (2.6)$$

/15...

The test statistic K is the square of the absolute maximum of a set of linear transformations of the V_i, and its cumulative distribution can be found by integrating the density function $h(\{v_i\} \mid \{f_i\})$ over an appropriate region. Thus, the cumulative null asymptotic permutation distribution of K is

$$D(k) = \text{Prob}(K \leq k)$$

$$= \int_R \cdots \int h(\{v_i\} \mid \{f_i\}) \, d\{v_i\} \qquad i = 1, 2, \ldots, (c-1)$$

$$\ldots\ldots\ldots\ldots\ldots\ldots \quad (2.7)$$

where the (c-1)-dimensional region of integration R is a simplex bounded by hyperplanes which are linear combinations of the V_i. This region of integration depends in general on the f_i, k and on the type of the predictor: whether it is monotonic, free or cyclic. Note that the asymptotic distribution of K does not depend on the sample size, nor on the absolute number of observations in each category, but on the number of categories and the proportion of the sample within each category.

It follows immediately that if there are only two categories (ie. c = 2) then $h(\{v_i\} \mid \{f_i\}) \sim N(0,1)$ and

$$K = M_1 N(\bar{X}_1 - \bar{\bar{X}})^2 / M_2 S^2$$

$$= (M_1/(N-M_1))N(\bar{X}_1 - \bar{\bar{X}})^2/S^2$$

$$= N(f_1/(1-f_1))(\bar{X}_1 - \bar{\bar{X}})^2/S^2$$

$$= (N^{\frac{1}{2}} G_1 (\bar{X}_1 - \bar{\bar{X}})/S)^2$$

$$= V_1^2 \sim \chi_1^2 \qquad \text{since } V_1 \sim N(0,1)$$

/16...

This result has been mentioned earlier in this chapter as being a generalization of the work by MacNaughton-Smith (1963).

Note that this result when $c = 2$ is independent of how the observations are allocated to the two categories (ie. independent of the f_i). Obviously, the distribution of K for $c > 2$ will be stochastically larger than this case as the number of categories in the predictor variable is increased. Thus the critical values from the χ^2_1 distribution form a lower bound for the critical values of K for any number of categories.

2.2 Distribution in the monotonic case.

Since a c-category monotonic predictor can only be split in c-1 ways, it is convenient to retransform the V_i to correspond to the K_i as defined in section 2.1. This will also have the edifying effect of simplifying the region of integration.

Let $W_i = (G_{1,2,\ldots,i}/(N_1+N_2+\ldots+N_i)) \sum_{j=1}^{i} N_j V_j/G_j$

$i = 1, 2, \ldots, c-1$

where $G_{1,2,\ldots,i} = \left((f_1+f_2+\ldots+f_i)/(1-f_1-f_2-\ldots-f_i) \right)^{\frac{1}{2}}$

and $G_i = \left(f_i/(1-f_i) \right)^{\frac{1}{2}}$ as before

yielding $q(\{w_i\} \mid \{f_i\}) \sim N(0,\Sigma_w)$

where Σ_w has covariance term

/17...

$$\sigma_w(i,j) = \mathrm{Cov}\left((G_{1,2,\ldots,i}/(N_1+N_2+\ldots+N_i)) \sum_{k=1}^{i} N_k V_k/G_k \right.,$$

$$\left. (G_{1,2,\ldots,j}/(N_1+N_2+\ldots+N_j)) \sum_{m=1}^{j} N_m V_m/G_m \right)$$

$$= G_{1,2,\ldots,i} G_{1,2,\ldots,j} / (N_1+N_2+\ldots N_i)(N_1+N_2+\ldots+N_j)$$

$$\times \mathrm{Cov}\left(\sum_{k=1}^{i} N_k V_k/G_k \, , \sum_{m=1}^{j} N_m V_m/G_m \right)$$

If $i \leq j$ then

$$\sigma_w(i,j) = G_{1,2,\ldots,i} G_{1,2,\ldots,j} / (N_1+N_2+\ldots N_i)(N_1+N_2+\ldots+N_j)$$

$$\times \mathrm{Cov}\left(\sum_{k=1}^{i} N_k^2 V_k^2/G_k^2 + \sum_{k} \sum_{\substack{m \\ k\neq m}} N_k N_m V_k V_m/G_k G_m \right)$$

$$= G_{1,2,\ldots,i} G_{1,2,\ldots,j} / (N_1+N_2+\ldots+N_i)(N_1+N_2+\ldots+N_j)$$

$$\times \left(\sum_{k=1}^{i} N_k(N-N_k) - \sum_{k} \sum_{\substack{m \\ k\neq m}} N_k N_m \right)$$

from 2.5 and 2.6

$$= G_{1,2,\ldots,i} \, G_{1,2,\ldots,j} / (N_1+N_2+\ldots+N_i)(N_1+N_2+\ldots+N_j)$$

$$\times \left(N \sum_{k=1}^{i} N_k - \sum_{k=1}^{i} \sum_{m=1}^{j} N_k N_m \right)$$

$$= G_{1,2,\ldots,i} ((N_1+N_2+\ldots+N_j)/(N-N_1-N_2-\ldots-N_j))^{\frac{1}{2}}$$

$$\times (N - \sum_{m=1}^{j} N_m) / (N_1+N_2+\ldots+N_j)$$

ie. $\sigma_w(i,j) = G_{1,2,\ldots,i} / G_{1,2,\ldots,j}$ for $i \leq j$

. . . . . (2.8)

/18...

and in particular for $i = j$ the variance terms are all unity.

Now

$$(N_1+N_2+\ldots+N_i)^{-1}\sum_{j=1}^{i} N_j V_j/G_j = (N_1+N_2+\ldots+N_i)^{-1}\sum_{j=1}^{i} N_j N^{\frac{1}{2}} G_j(\bar{X}_j - \bar{X})/G_j S$$

$$=(N^{\frac{1}{2}}/S)\sum_{j=1}^{i} N_j(\bar{X}_j - \bar{X})/(N_1+N_2+\ldots+N_i)$$

$$= N^{\frac{1}{2}}(\bar{X}_{(i)} - \bar{X})/S \qquad \ldots\ldots\ldots (2.9)$$

where $\bar{X}_{(i)}$ is the mean of the first i categories.

Hence

$$W_i^2 = G_{1,2,\ldots,i}^2 \left(\sum_{j=1}^{i} N_j V_j/G_j(N_1+N_2+\ldots+N_i)\right)^2$$

$$= G_{1,2,\ldots,i}^2 \; N(\bar{X}_{(i)} - \bar{X})^2/S^2 \quad \text{from 2.9}$$

$$= \left((f_1+f_2+\ldots+f_i)/(1-f_1-f_2-\ldots-f_i) \right) N(\bar{X}_{(i)} - \bar{X})^2/S^2$$

$$= \left((N_1+N_2+\ldots+N_i)/(N-N_1-N_2-\ldots-N_i) \right) N(\bar{X}_{(i)} - \bar{X})^2/S^2$$

$$= \left(N_{(i)}/(N-N_{(i)}) \right) N(\bar{X}_{(i)} - \bar{X})^2/S^2$$

$$\ldots\ldots\ldots\ldots (2.10)$$

where $N_{(i)}$ is the total number of observations in the first i categories, and so $W_i^2 = K_i$, the statistic obtained by splitting the c-category monotonic predictor between categories i and $i+1$. It follows that the condition $K \leq k$ is equivalent to $|W_i| \leq \sqrt{k}$ for all i, so 2.7 becomes

$$D(k) = \text{Prob}(K \leq k) = \int_{-\sqrt{k}}^{\sqrt{k}} \int_{-\sqrt{k}}^{\sqrt{k}} \cdots \int_{-\sqrt{k}}^{\sqrt{k}} q(\{w_i\} \,|\, \{f_i\}) \; d\{w_i\}$$

$$i = 1, 2, \ldots, (c-1)$$

$$\ldots\ldots\ldots\ldots (2.11)$$

/19...

That is, the region of integration becomes a (c-1)-dimensional hypercube centred at the origin with sides of length $2\sqrt{k}$. The evaluation of this integral is postponed to chapter 3.

2.4 Distribution in free case.

In this case, returning to 2.7

$$D(k) = \text{Prob}(K<k) = \int\int \cdots \int_R h(\{v_i\}|\{f_i\})\, d\{v_i\} \qquad i = 1, 2, \ldots, (c-1)$$

the region of integration R is a (c-1)-dimensional polygon bounded by the $(2^c - 2)$ hyperplanes obtained by considering all possible subsets of the c categories (excluding both the null set and the universal set). Any combination of categories used to define a subset, yields a test statistic K_s that is the square of some linear combination of the V_i. For example, suppose the three categories i, j and k are combined.

Let $N_{(i,j,k)} = N_i+N_j+N_k$ be the total number of observations in these three categories

and $\bar{X}_{(i,j,k)} = (N_i\bar{X}_i+N_j\bar{X}_j+N_k\bar{X}_k)/N_{(i,j,k)}$ be the mean of these three categories

and $H^2_{i,j,k} = \left((f_i+f_j+f_k)(1-f_i-f_j-f_k) \right)^{-1}$

so that $V_i/H_i = N^{\frac{1}{2}}G_i(\bar{X}_i - \bar{X})/H_iS$

$= N^{\frac{1}{2}}f_i(\bar{X}_i - \bar{X})/S$

/20...

Then the test statistic obtained from this grouping is

$$N_{(i,j,k)}N(\bar{X}_{(i,j,k)}-\bar{X})^2/S^2(N-N_{(i,j,k)})$$

$$= (N/S^2)\ ((N_i+N_j+N_k)/(N-N_i-N_j-N_k))\ (\bar{X}_{(i,j,k)}-\bar{X})^2$$

$$= (N/S^2)\ ((f_i+f_j+f_k)/(1-f_i-f_j-f_k))$$

$$\times\ (\ (f_i\bar{X}_i+f_j\bar{X}_j+f_k\bar{X}_k - (f_i+f_j+f_k)\bar{X})/(f_i+f_j+f_k)\ )^2$$

$$= NH^2_{i,j,k}\ (\ f_i(\bar{X}_i-\bar{X})+f_j(\bar{X}_j-\bar{X})+f_k(\bar{X}_k-\bar{X})\ )^2\ /\ S^2$$

$$= H^2_{i,j,k}\ (\ V_i/H_i + V_j/H_j + V_k/H_k\ )^2$$

In general the (2^c-2) hyperplanes of the region of integration R in 2.7 can be written as

$$-\sqrt{k} \leq v_i \leq \sqrt{k} \qquad i = 1, 2, \ldots, (c-1)$$

$$-\sqrt{k} \leq H_{i,j}(v_i/H_i+v_j/H_j) \leq \sqrt{k} \qquad i,j = 1, 2, \ldots, (c-1) \quad i<j$$

$$-\sqrt{k} \leq H_{i,j,m}(v_i/H_i+v_j/H_j+v_m/H_m) \leq \sqrt{k} \quad i,j,m=1,2,\ldots,(c-1) \quad i<j<m$$

$$\vdots$$

$$-\sqrt{k} \leq H_{1,2,\ldots,(c-1)}(v_1/H_1+v_2/H_2+\ldots+v_{c-1}/H_{c-1}) \leq \sqrt{k}$$

$$\ldots\ldots\ldots\ldots\ldots\ldots \quad (2.12)$$

where the general term

$$H_{i,j,\ldots,p} = (\ (f_i+f_j+\ldots+f_p)(1-f_i-f_j-\ldots-f_p)\ )^{-\frac{1}{2}}$$

If each category contains the same number of cases then $f_1 = f_2 = \ldots = f_c = 1/c$, and the region of integration is defined by the hyperplanes

$$-(\ ks(c-s)/(c-1)\ )^{\frac{1}{2}} \leq v_{i(1)}+v_{i(2)}+\ldots+v_{i(s)} \leq (\ ks(c-s)/(c-1)\ )^{\frac{1}{2}}$$

for $s, i(1), i(2), \ldots, i(s) = 1, 2, \ldots, (c-1)$

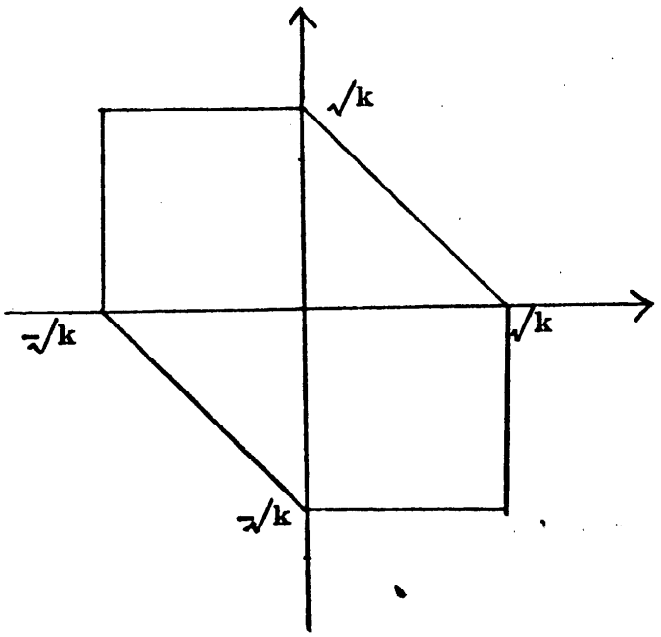
and $i(1) < i(2) < \ldots < i(s)$

$\ldots\ldots\ldots\ldots$ (2.13)

/21

The appearance of these intersecting hyperplanes is like a hypercube with various edges chamfered and certain corners lopped off. For example, if there is an equal proportion of observations within each category, then the two dimensional region of integration for the three free category case appears as in figure 2.1 below.

Figure 2.1. Two dimensional region of integration for three free categories with equal numbers of observations within each category.



and the integral may be written explicitly as

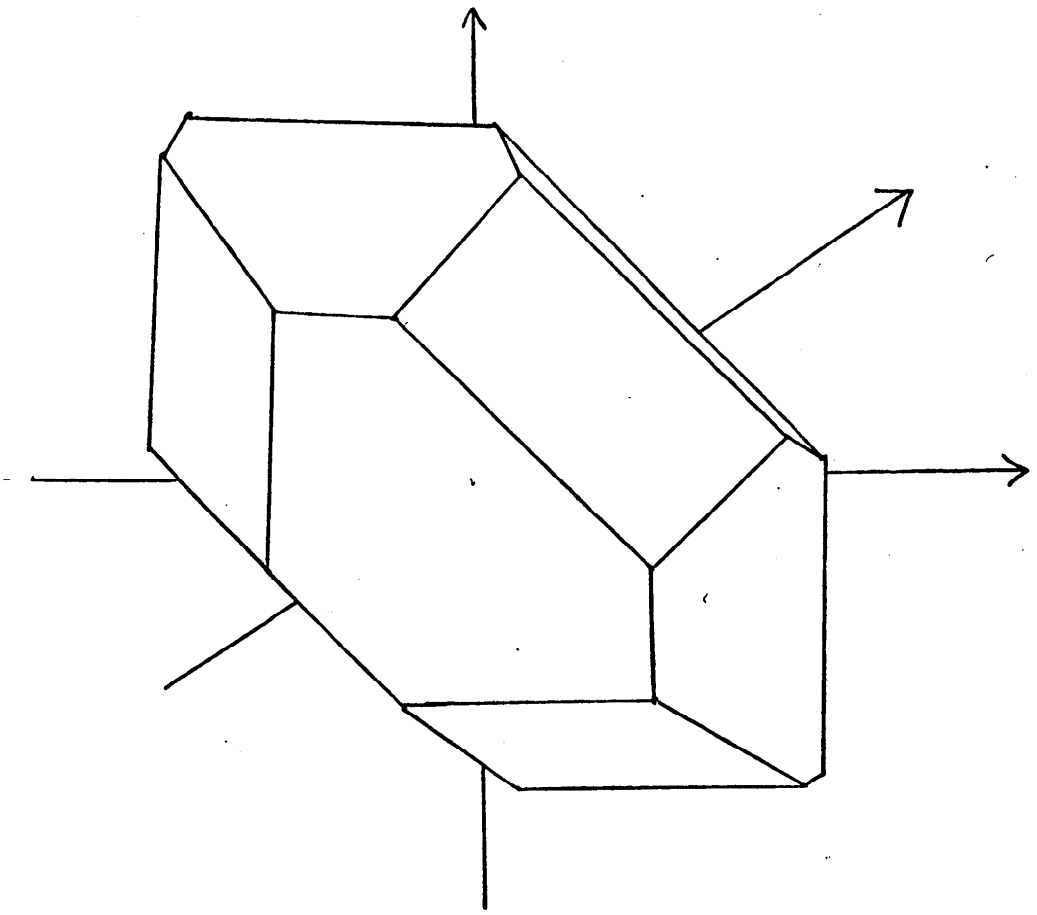
$$\int_{-\sqrt{k}}^{\sqrt{k}} \int_{\max(-\sqrt{k},\, -v_2-\sqrt{k})}^{\min(\sqrt{k},\, -v_2+\sqrt{k})} h(\{v_1,v_2\}\,|\,\{f_1,f_2\})\, dv_1\, dv_2 \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots (2.14)$$

in this case of equal proportions of observations within each category.

/22...

In the four category case the region of integration appears as in figure 2.2 below.

Figure 2.2. Three dimensional region of integration for four free categories with equal numbers of observations within each category.



and the integral 2.7 may be written explicitly as

$$\int_{-\sqrt{k}}^{\sqrt{k}} \int_{\max(-\sqrt{k},\, -v_3-\sqrt{(ek)})}^{\min(\sqrt{k},\, -v_3+\sqrt{(ek)})} h(\{v_1,v_2,v_3\} \mid \{f_1,f_2,f_3\})\, dv_1\, dv_2\, dv_3 \quad \ldots\ldots\ldots\ldots (2.15)$$

the innermost integral ranging from

$$\max\left(-\sqrt{k},\ -v_3-\sqrt{(ek)},\ -v_2-\sqrt{(ek)},\ -v_2-v_3-\sqrt{k} \right)$$

to

$$\min\left(\sqrt{k},\ -v_3+\sqrt{(ek)},\ -v_2+\sqrt{(ek)},\ -v_2-v_3+\sqrt{k} \right)$$

/23...

and where in this case of equal proportions of observations within each category

$e = s(c-s)/(c-1) = 4/3$

Again, the evaluation of these integrals is postponed to chapter 3.

2.5 Note on distribution in cyclic case.

Since the cyclic case does not appear to be applied, remarks on this topic will be brief and included only for the sake of completeness.

The complexity involved here is between that of the monotonic case and that of the free case. Following the argument used to establish the region of integration in the free case, it is seen that a small subset of the $(2^c - 2)$ hyperplanes define the required region in this case. Since the chosen categories determining a subgroup must be contiguous (with "wraparound") the required planes are

$$-\sqrt{k} \leq v_i \leq \sqrt{k} \qquad i = 1, 2, \ldots, (c-1)$$

$$-\sqrt{k} \leq H_{i,i+1}(v_i/H_i + v_{i+1}/H_{i+1}) \leq \sqrt{k} \qquad i = 1, 2, \ldots, (c-2)$$

$$-\sqrt{k} \leq H_{i,i+1,i+2}(v_i/H_i + v_{i+1}/H_{i+1} + v_{i+2}/H_{i+2}) \leq \sqrt{k} \qquad i=1,2,\ldots,(c-3)$$

$$\vdots$$

$$-\sqrt{k} \leq H_{1,2,\ldots,(c-1)}(v_1/H_1 + v_2/H_2 + \ldots + v_{c-1}/H_{c-1}) \leq \sqrt{k}$$

. . . . . (2.16)

/24...

CHAPTER 3. Computer Evaluation of Distribution.

This chapter concerns the computer evaluation of the asymptotic distribution of K from the integral expressions in chapter 2. The distribution for a c-category predictor involves a (c-1)-dimensional integral which cannot be expressed in closed form in general. Its computer evaluation is considered in two parts: firstly where there are only a few categories and numerical quadrature can be employed, and the second where there are many categories and a Monte Carlo approach is more suitable.

3.1 Evaluation by numerical integration.

When there are only two categories the distribution of K follows the familiar chi-square distribution with one degree of freedom. For the purpose of providing figures in our tables, the integral was evaluated by using the IBM supplied subroutine ERF which is available both from FORTRAN and PL/1, see IBM(1968). This routine that evaluates the error function is defined by

$$\mathrm{erf}(x) = (2/\sqrt{\pi}) \int_0^x \exp(-t^2)\, dt \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.1)$$

A simple transformation shows that if $Z \sim N(0,1)$ then

$$\mathrm{Prob}(-z<Z<z) = (1/\sqrt{(2\pi)}) \int_{-z}^{z} \exp(-\tfrac{1}{2}t^2)\, dt = \mathrm{erf}(z/\sqrt{2}) \qquad \ldots\ldots\ldots\ldots\ldots\ldots (3.2)$$

using the symmetry of the Normal distribution about zero.

Since $K \sim \chi^2_1$ (the square of a standard normal deviate) in the two category case, we have

$$D(k) = \text{Prob}(K \leq k) = \text{erf}(\sqrt{(\tfrac{1}{2}k)}) \qquad \ldots\ldots\ldots\ldots (3.3)$$

When there are only three categories, ie. $c = 3$, the bivariate normal integral results which can be evaluated by a number of methods, see Gupta (1963). The integral obtained for a three category monotonic predictor can be found in tables by Krishnaiah and Armitage (1965) by referring to the tables for the maximum of two correlated chi-square variates. This paper also gives the upper percentage points for these distributions.

In the monotonic case the dimension of the integral can be halved for numerical integration by the method of Curnow and Dunnett (1962) who give reduction formulae for evaluating multivariate normal integrals when the correlation term in position i,j $(i<j)$ is of the form γ_i/γ_j with $|\gamma_i| \leq |\gamma_j|$. Now the monotonic case requires the integration over a hypercube centred at the origin of a multivariate normal with correlation matrix having term $G_{1,2,..,i}/G_{1,2,..,j}$ in position (i,j) for $i<j$ (see 2.8). The numerator and denominator of this covariance term are both positive and further

$$G_{1,2,..,i}/G_{1,2,..,j} = \left(\frac{f_1+f_2+\ldots+f_i}{1-f_1-f_2-\ldots-f_i} \cdot \frac{1-f_1-f_2-\ldots-f_j}{f_1+f_2+\ldots+f_j} \right)^{\frac{1}{2}}$$

$$= \left(\frac{f_1+f_2+\ldots+f_i}{f_1+f_2+..+f_i+..+f_j} \cdot \frac{1-f_1-f_2-..-f_i-..-f_j}{1-f_1-f_2-\ldots-f_i} \right)^{\frac{1}{2}}$$

< 1 for $i<j$ since each $f_i>0$ and $\Sigma f_i=1$

/26...

so that the method of Curnow and Dunnett is applicable. In practice the use of this method becomes tedious when the number of dimensions is not small.

A general computer program for evaluating multivariate normal integrals was obtained from Dr. R. C. Milton who has described it in Milton (1972). Because of the symmetry involved, only half the integral need be evaluated by reducing the range of the outermost integral from $(-\sqrt{k}, \sqrt{k})$ to $(0, \sqrt{k})$. This multidimensional quadrature program was used to evaluate the distribution of K for a monotonic predictor with 3, 4 or 5 categories, and for a free predictor with 3 or 4 categories.

For larger numbers of categories direct numerical integration was found to involve excessive use of computer time, so Monte Carlo methods were employed.

3.2 Evaluation by Monte Carlo methods.

In both the monotonic and free cases the integrals were evaluated by the Monte Carlo technique described below, for the number of categories c ranging from 4 to 10 and for integral values of k from 4 upwards. The results for $c = 4$ and 5 were checked against the figures obtained by numerical integration as in section 3.1 above.

/27...

The following four subsections describe the various stages of the Monte Carlo evaluation. The first subsection discusses the analytic evaluation of part of the region of integration. The second subsection discusses the stratification of that part of the integral that has not been evaluated analytically. The third subsection considers the generation of pseudorandom normal deviates which are required in the Monte Carlo evaluation. These three subsections are presented as general techniques useful in evaluating multivariate normal integrals and are not restricted to the particular integrals of this thesis. The last subsection applies these results to the integrals of chapter 2 and introduces a further refinement of the Monte Carlo estimates.

3.2.1 Partial analytic evaluation of multivariate normal integrals.

In this section the general problem of the integration of a multivariate normal distribution over a closed region by Monte Carlo methods is considered. Many statistical problems require this evaluation especially near the tails of the distribution in order to obtain critical values. For example, outlier theory requires the absolute maximum of a set of normal variables, which yield a hypercube centered at the origin as the region of integration as in our monotonic case. Other problems involving quadratic forms

/28...

for example require integration within hyperspheres or hyperellipsoids, and so on.

The problem without loss of generality is to evaluate

$$\int_{b_n}^{u_n} \int_{b_{n-1}}^{u_{n-1}} \cdots \int_{b_i}^{u_i} \cdots \int_{b_1}^{u_1} \phi(\{x_i\})\, d\{x_i\} \qquad i = 1 \text{ to } n$$

where $\phi(\{x_i\})$ is multivariate normal with zero mean vector and covariance matrix equal to the identity matrix. The limits of integration b_i and u_i $(i<n)$ are some function of the x_j $(j>i)$, and b_n and u_n are constants.

Integration of a multivariate normal with arbitrary mean vector and covariance matrix can be easily turned by a linear transformation into the above form, usually resulting in a modification to the region of integration. For example, the change in a spherical or elliptical region is to another elliptical region, and hyperplanes remain hyperplanes though with a change in orientation. The region can often be further simplified by a rotation of the axes which leaves unchanged the spherical normal integral.

The idea considered here is the well known concept of evaluating analytically as much as possible of the unknown region, see for example Hammersley and Handscomb (1964). The basic method is to place a hypercube with edges parallel to the axes inside the region of integration. Its

/29...

probability content is easily determined and the eventual Monte Carlo evaluation will then be confined to the exterior of this analytic region.

While a suitably chosen hyperrectangle will in general have more probability content than a hypercube, the problem of finding the optimum hyperrectangle is a difficult one and beyond the scope of this thesis. Also other shaped regions of integration of known probability content could be used but difficulty arises in sampling points outside them, and sometimes of ascertaining whether they lie completely within the region of integration.

If the hypercube has sides of length $2a$ then its probability content is

$$(\Phi(a) - \Phi(-a))^d = p^d \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (3.4)$$

where d is the number of dimensions of the region, and $\Phi(x)$ is the cumulative standard normal distribution function. Ideally p^d should be made as large as possible subject to the condition that the hypercube is completely enclosed by the region of integration. To avoid confusion this hypercube will be termed the "interior-hypercube" in the sequel.

Finding the largest interior-hypercube may be done analytically, or iteratively by computer. In the latter case a convex region of integration simplifies matters since then it is only necessary to check that the corners of the interior-hypercube lie within the region of integration. Also

it may be noted that while the maximum sized interior-hypercube is theoretically preferable, no bias is introduced by using a smaller interior-hypercube.

The gains realised by this method were compared for various integrals by examining the probability content of the interior-hypercube for different critical values. The optimum interior-hypercube is easily found analytically for ellipsoidal regions of integration. Remembering that the transformation to the integration of a spherical normal from the non-zero correlation case merely transforms ellipsoids to other ellipsoids, the axes should be further rotated if necessary so that the axes of integration and the major axes of the ellipsoids coincide. The largest interior-hypercube is obtained when its corners lie on the surface of the ellipsoid. Hence, in general the largest interior-hypercube inside

$$a_1X_1^2 + a_2X_2^2 + \ldots + a_nX_n^2 = r$$

cuts each axis at $\pm\sqrt{r/(a_1+a_2+\ldots+a_n)}$

Tables of the critical values for this region of integration are given by Owen (1962). The following tables 3.1 and 3.2 give the probability content of the interior-hypercube centred at the origin within various two and three-dimensional ellipses which correspond to the 90% and 95% critical values.

/31...

Table 3.1. Probability content of the largest standard interior-hypercube within ellipse $a_1X_1^2 + a_2X_2^2 = r$

		Probability content of ellipse	
a_1	a_2	.9	.95
.50	.50	.76	.84
.75	.25	.77	.86
.90	.10	.79	.89
.99	.01	.81	.90

Table 3.2. Probability content of the largest standard interior-hypercube within ellipsoid $a_1X_1^2 + a_2X_2^2 + a_3X_3^2 = r$

			Probability content of ellipsoid	
a_1	a_2	a_3	.9	.95
.33	.33	.33	.62	.71
.50	.40	.10	.64	.75
.80	.10	.10	.68	.81
.90	.05	.05	.71	.84

As can be seen quite a large proportion of the integral is contained within the interior-hypercube. The minimum content occurs where the region is a circle or sphere, ie. when $a_1=a_2=\ldots$, and the proportion decreases as the number of dimensions increase. It is of course possible that an interior-hyperrectangle can be found with larger probability content - a possible subject for future research.

/32...

The problem of finding the largest interior-hypercube where the region of integration is a hypercube in a different space is more difficult. Let the correlation matrix of the integrand be

$$R = C'C$$

where the matrix $C = (c_{ij})$ is some square root of the symmetric matrix R. By standard transformation theory the interior-hypercube with corners at $(\pm a, \pm a, \ldots, \pm a)$ will in the space of the integrand be a parallelepiped with corners generated by

$$a \sum_i \pm c_{ij}$$

considering all possible combinations of +'s and -'s, so that the largest value on a dimension is given by

$$a \left(\max_j \sum_i |c_{ij}| \right)$$

If the region of integration is given by $(\pm b, \pm b, \ldots, \pm b)$ in the non-spherical space of the integrand, then the maximum sized interior-hypercube for this orientation has corners at

$$\pm b \Big/ \left(\max_j \sum_i |c_{ij}| \right) \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (3.5)$$

A convenient choice for C is the so-called Choleski square root of R, that is a triangular matrix C such that $C'C = R$. This factorization is readily available under the name MFSD in IBM's scientific subroutine package, IBM (1970).

/33...

The efficacy of this method is examined for an equicorrelated integrand since tables of the critical values have been published by Krishnaiah and Armitage (1965). The following tables 3.3, 3.4 and 3.5 give the probability content of the interior-hypercube in spherical normal space centred at the origin that lies completely within the hypercubic region of integration for equicorrelations of 0.1, 0.3 and 0.8 in 2, 4, 6, 8 and 10 dimensions.

Table 3.3. Probability content of standard interior-hypercube within a hypercube where all correlations are 0.1

Content of hypercube	Number of dimensions 2	4	6	8	10
.90	.86	.72	.57	.42	.29
.95	.92	.82	.69	.54	.40
.99	.98	.94	.86	.76	.64

Table 3.4. Probability content of standard interior-hypercube within a hypercube where all correlations are 0.3

Content of hypercube	Number of dimensions 2	4	6	8	10
.90	.77	.47	.25	.12	.06
.95	.85	.59	.35	.19	.10
.99	.95	.78	.57	.37	.23

/34...

Table 3.5. Probability content of standard interior-hypercube within a hypercube where all correlations are 0.8

Content of hypercube	Number of dimensions 2	4	6	8	10
.90	.66	.32	.15	.06	.03
.95	.77	.45	.24	.12	.06
.99	.90	.67	.46	.30	.19

As can be seen, the improvement due to the partial analytic evaluation decreases with increasing number of dimensions and with increase of the common intercorrelation. For less than six dimensions the gain is quite appreciable but appears not so impressive for as many as ten dimensions. These tables do not reflect the optimum orientation of the interior-hypercube since the choice of C in 3.5 is not unique, other roots being related by some rotation. The choice of the Choleski (triangular) square-root minimizes the computation required to transform points from the spherical normal space to the space of the integrand, other roots in general requiring about double the number of computations. The gain from decreasing the variance of the Monte Carlo estimate, E, say, is approximately linear in the probability content of the interior-hypercube since E is of the form

$$E = p^d + (1-p^d)\hat{q}$$

where p^d is the content of the interior-hypercube

/35...

as defined in 3.4 and $\hat{q}$ is the estimate of the conditional probability content of the region of integration outside the interior-hypercube. If M points are sampled outside the interior-hypercube then

$$\text{Var}(E) \doteq (1-p^d)^2 \, \hat{q}(1-\hat{q}) \, / \, M$$

and if the actual percentage point is t then

$$\hat{q} \doteq (t-p^d) \, / \, (1-p^d)$$

so

$$\text{Var}(E) \doteq (1-p^d)^2 \, \frac{(t-p^d)(1-p^d-(t-p^d))}{M\,(1-p^d)^2}$$

$$= (t-p^d)\,(1-t) \, / \, M \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (3.6)$$

which is linear in p^d.

The estimate of the variance of E is further complicated by the stratified sampling that will be introduced to select points outside the interior-hypercube. It is not clear at this stage that increasing the probability content of the interior-hypercube by change in orientation will necessarily increase or even leave unchanged the precision of the sampling method. A further disadvantage of seeking a better orientation for the interior-hypercube is that at this stage of the development it is not easy to obtain the optimally orientated interior-hypercube. The mathematics of this optimization requires the minimization of

$$\max_j \sum_i |c_{ij}| \quad \text{subject to} \quad C'C = R, \text{ from 3.5.}$$

/36...

This is a linear programming problem with quadratic constraints, and has no general explicit solution. An attempt was made to solve this iteratively by optimally rotating two axes of C at a time, either trying each pair in turn, or first examining all pairs of axes before performing the rotation. When there where more than three dimensions, this process stabilized at a local minimum which unfortunately was not the global optimum, a better solution being found by inspection. In order to smooth the objective function, experiments were made with minimizing functions of the form

$$\sum_j \left(\sum_i |c_{ij}|\right)^r$$

for various values of r. Note that as $r \to \infty$ this latter objective function gives equivalent results to minimizing the original function. Again only local minima were attained. Note in passing that this problem is superficially similar to Varimax rotation in Factor Analysis, see for example Lawley and Maxwell (1971). However Varimax rotation is considerably simplified by the lack of modulus signs in the criterion. Table 3.6 compares the gains for the interior-hypercube of largest probability content found with those of the Choleski orientation in ten dimensions with an equicorrelation of 0.3.

/37...

Table 3.6. Comparison of the probability content of the best orientated hypercube found with the orientation obtained by Choleski's square root, in ten dimensions and all correlations equal to 0.3.

Content of hypercube, t	Choleski transformation	Best transformation found.
.90	.06	.23
.95	.10	.34
.99	.23	.58

The overall gain is of the order of three and decreases as the critical value increases. Substituting these values in 3.6 the leading multiplier decreases from .84 to .67 when t=.9 and from .86 to .41 when t=.99. This reduction of projected variance is not very encouraging. More research and experimentation that is beyond the scope of this thesis is required both to determine the optimum orientation of the interior-hypercube and a computationally efficient method of finding this orientation.

The transformation provided by the Choleski square root requires about half the computational effort demanded by other non-triangular transformations. Since the variance is inversely proportional to the sample size, a reduction to at least half the variance is needed before a non-triangular transformation matrix can be considered viable. At this stage it appears more cost effective to accept the definite reduction in computational effort provided by the Choleski square root, than the usually non-compensating gain of a better orientation (in terms of the probability content) of the interior-hypercube.

Again, the interior-hypercube may not be the optimum shape to place inside the region of integration, but some inte-

rior-hyperrectangle may provide even more probability content. Attention has been given to rectangular figures because of the comparative ease of sampling outside them and which also allows an added sophistication that is now considered.

3.2.2 Stratification outside the interior-hypercube.

The probability content of the interior-hypercube is evaluated analytically. Outside this region points must be sampled for the Monte Carlo evaluation. It is well known, see for example Cochran (1963), that stratified sampling is usually more accurate than simple random sampling, and is never less accurate if each stratum is sampled proportionately to its size. In our particular example the stratification that is proposed will have definite gains over simple random sampling since the strata will not in general be homogeneous. The use of the interior-hypercube with probability content as given in 3.4

$$(\Phi(a) - \Phi(-a))^d = p^d$$

where d is the number of dimensions and the interior-hypercube has sides of length $2a$ lends itself to the formation of d non overlapping strata as defined in the following table 3.7 where the fraction of the sample that yields a selection proportional to the size of the stratum is also given.

/39...

Table 3.7. Proportional stratification of region outside interior-hypercube.

Take proportion	From the region
$(1-p)/(1-p^d)$	$-\infty<x_1,x_2,\ldots,x_{d-1}<\infty,\ \lvert x_d\rvert>a$
$p(1-p)/(1-p^d)$	$-\infty<x_1,x_2,\ldots,x_{d-2}<\infty,\ \lvert x_{d-1}\rvert>a,\ -a\leq x_d<a$
$p^2(1-p)/(1-p^d)$	$-\infty<x_1,x_2,\ldots,x_{d-3}<\infty,\ \lvert x_{d-2}\rvert>a,\ -a\leq x_{d-1},x_d\leq a$
⋮	⋮ ⋮ ⋮ ⋮ ⋮ ⋮ ⋮ ⋮ ⋮ ⋮ ⋮ ⋮ ⋮ ⋮ ⋮
$p^{d-1}(1-p)/(1-p^d)$	$\lvert x_1\rvert>a,\ -a<x_2,x_3,\ldots,x_d\leq a$

In practice a sample vector $(x_1,x_2,\ldots,x_d)$ was obtained by sampling each element independently from a standard normal distribution, with the restrictions as set out above, and then these vectors were transformed into the appropriate space of the integrand. There are a number of Monte Carlo methods that could be employed at this stage, see Hammersley and Handscomb (1964), but the majority that required evaluation of the multinormal density were rejected as being too time consuming compared with the actual generation of the random normal vectors. The hit-or-miss method given in the above reference was employed and found to give sufficient accuracy (discussed in 3.4 below). This method requires a large quantity of random normal numbers, many of them constrained to a particular range. A fast method for their generation is discussed in the next subsection.

/40...

3.2.3 Generation of pseudo-random normal deviates.

A number of methods for generating random normal numbers have appeared in the literature. A summary of the better ones is given in Knuth (1969), who shows the faster methods to be of the class typified by Marsaglia, MacLaren and Bray (1964). The main disadvantage of these methods in our application is that they cannot be modified to handle a normal deviate in a particular range except by rejecting "unsuitable" numbers. The conceptually obvious way of handling this problem, is to firstly generate a uniform random number U within the appropriate range and obtain the random normal deviate by the inverse transformation $\Phi^{-1}(U)$. This follows from the well known result (see for example Kendall and Stuart (1958), p.18) that any continuous cumulative distribution function is rectangularly distributed. Unfortunately, no closed form of this inverse transformation exists, and one must use an approximation. Two reasonable approximations are Burr (1967), (as given in Meyer (1969)),

$$X = \frac{((1-U)^{-1/6.158} - 1)^{1/4.874} - (U^{-1/6.158} - 1)^{1/4.874}}{.323968}$$

and the more accurate Hastings (1955),

$$X = \delta\left(W - \frac{2.515517 + .802853W + .010328W^2}{1 + 1.432788W + .189269W^2 + .001308W^3}\right)$$

where $\delta = 1$, $W = (-2\ln(U))^{\frac{1}{2}}$ if $0 < U \leq .5$

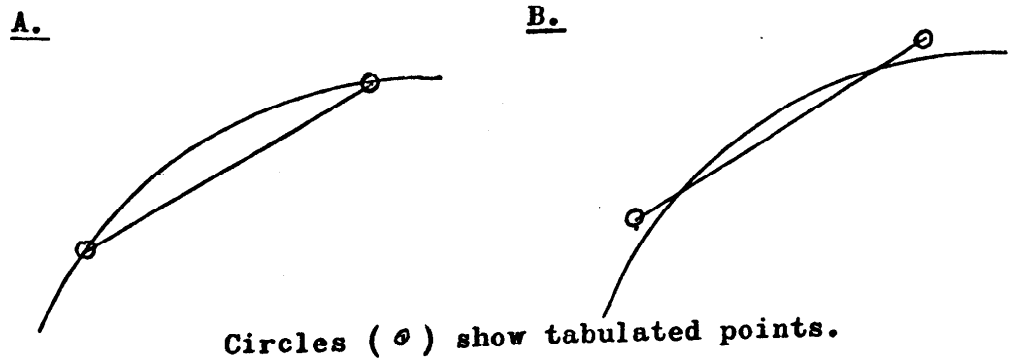
and $\delta = -1$, $W = (-2\ln(1-U))^{\frac{1}{2}}$ if $.5 < U \leq 1$

/41...

Both these methods are computationally inefficient since the first requires four exponentiations, and the second a logarithm and a square root. To avoid these time consuming operations as far as possible, much of the inversion was performed by table look-up. The symmetry of the normal distribution $\Phi^{-1}(U) = -\Phi^{-1}(1-U)$ was used to reduce the span of the tables to $U \leq .5$.

The first table gave $\Phi^{-1}(U)$ in the range $0 < U \leq .5$ in steps of .0005 (ie. 1001 points), and in this algorithm was only used when $U > .01$ which holds 98% of the time if the required random normal deviate is unrestricted. Since the graph of $X = \Phi^{-1}(U)$ is convex in this range of $0 < U \leq .5$ it meant that linear interpolation in the table would always underestimate X, see figure 3.1A below.

Figure 3.1 Linear interpolation in the function $X=\Phi^{-1}(U)$

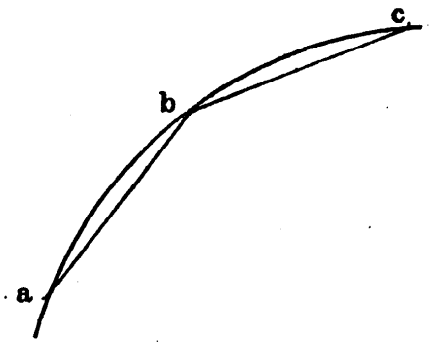


Circles (⊙) show tabulated points.

To counteract this, the tabulated points were increased to overestimate X at these points and reduce the underestimation within the range, see figure 3.1B above.

This was accomplished by a first approximation to the well-known minimax solution by averaging the error at the midpoint of the adjacent ranges (a,b) and (b,c) in figure 3.2 below. This correction was then applied to point b, and similarly for every other tabulated point.

Figure 3.2 Linear interpolation before correction is applied.



The maximum absolute error in the table for the range $.01 \leq U < .5$ is 5.3×10^{-5} attained when $U=.01$ ($X=-2.326295$ instead of $X=-2.326348$). This gives five significant figures which were also retained when X was less than one since the absolute error was of the order of 10^{-6} at this part of the table.

The second table gave $\Phi^{-1}(U)$ in the range $0<U<.01$ in steps of .00001 (ie. 1001 points) and in this algorithm was only used when $.0001<U<.01$. Again, correction was made to the table as in the first to prevent continual underestimation of the inverse transformation. The maximum absolute error in this table for the range $.0001<U<.01$ is 1.5×10^{-4} attained when $U=.0001$ ($X=-3.71887$ instead of $X=-3.71902$). Outside the range of the tables, ie. when

/43...

$U<.0001$, the Hasting's approximation as cited above with absolute maximum error 4.5×10^{-4} was used.

The selection of a suitable uniform random number generator for the particular computer available was conveniently simplified by the article of Whittlesey (1968) who compared a number of such generators on the IBM 360. His recommendations preferring the Tausworthe generator (Tausworthe, 1965) were followed especially as its speed was higher than that of others considered. An Assembler language (the lowest level language of the IBM 360, 370) routine was written incorporating the points mentioned above, and appears in Appendix A. This routine provided a vector of independent random normal deviates in the appropriate range as defined by Table 3.7.

3.2.4 Results of Monte Carlo evaluation.

The routine of section 3.2.3 was incorporated into a program to evaluate the distribution of K when all categories of the predictor had an equal allocation of observations. This integral $D_c(k)$ for c categories was estimated for integral values of k from 4 upwards ensuring that the highest value of k exceeded the 99% critical value, so that subsequent evaluation of the critical value would be by interpolation rather than extrapolation.
The Monte Carlo technique examined 25000 data points in the monotonic case when there were six or fewer categories,

/44...

and 10000 data points in all other cases. Besides the estimated value of the integral, a 95% confidence interval was also established for each estimate.

Knowing that the upper tail of the χ_1^2 distribution is approximately of exponential form, and that this distribution forms a lower limit to $D_c(k)$ for $c>2$ as mentioned at the conclusion of section 2.2, it was felt that $D_c(k)$ itself might also be approximately of exponential form so the Monte Carlo estimates were further refined by fitting regression lines to $\ln(1-\hat{D}_c(k))$ weighted by

$$(\ln(1 - \text{lower limit}) - \ln(1 - \text{upper limit}))^{-1}$$

which is a simple function of the range of the 95% confidence interval $\hat{D}_c(k)$ obtained earlier transformed into the logarithmic scale.

Experimentation revealed that in the monotonic case, the model

$$\ln(1-D_c(k)) = \alpha_c + \beta k \quad \text{(No. of categories } c = 5,6,..,10)$$

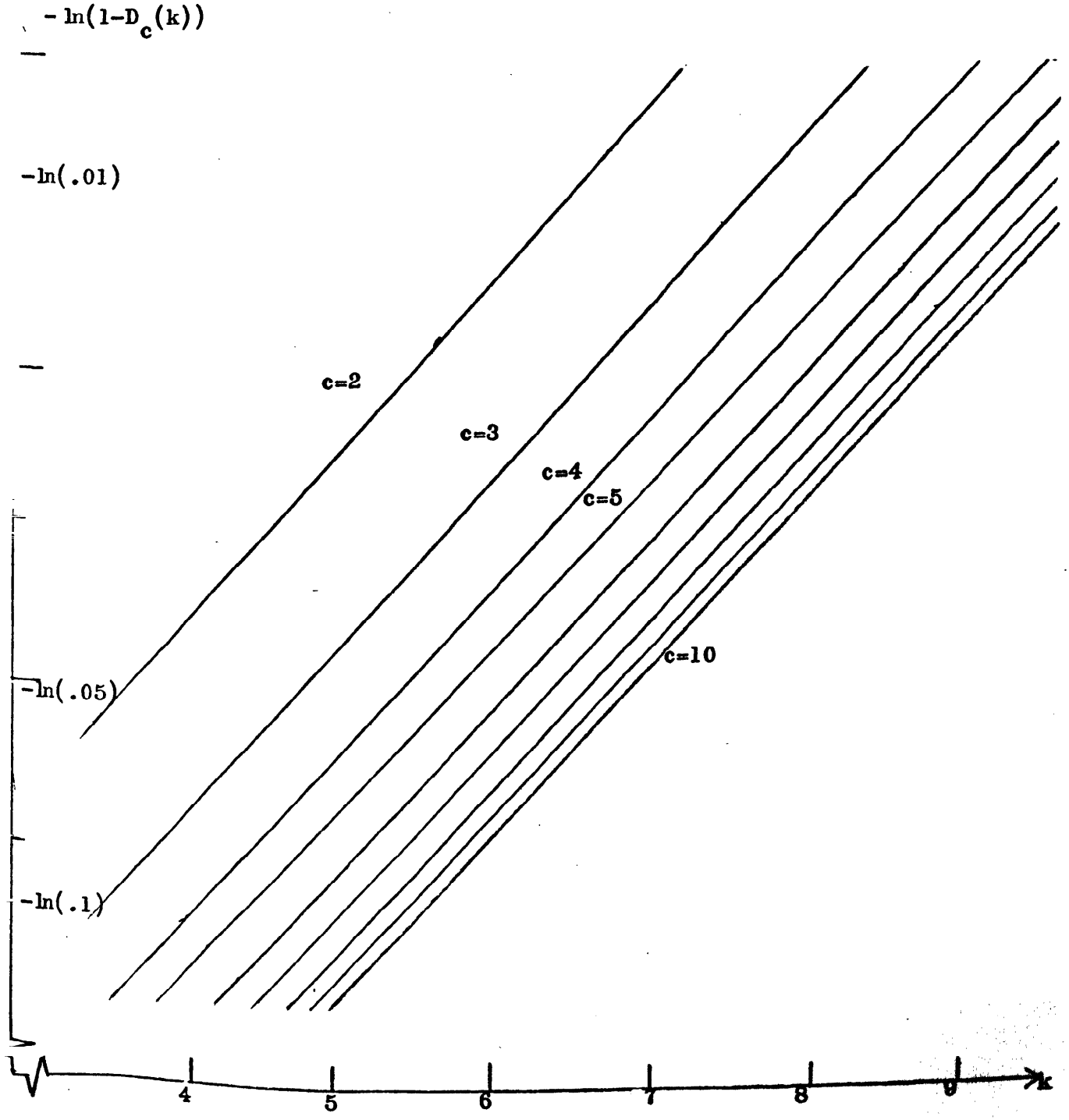
for $D_c(k) > .75$ gave an excellent fit with a multiple correlation of .99995. The estimate $\hat{\beta}=-.5303$ (the same for all numbers of categories c) had a standard error of .0017. The estimates of α_c and their standard errors are given in table 3.8, and these results are summarized in the graph of figure 3.3.

/45...

Table 3.8. The regression coefficients α_c in the monotonic case for the model $\ln(1-D_c(k))=\alpha_c+\beta k$ $\hat{\beta}=-.5303$, s.e.$(\hat{\beta})=.0017$

No. of categories c	5	6	7	8	9	10
$\hat{\alpha}_c$	.147	.308	.443	.552	.633	.687
St. error of $\hat{\alpha}_c$	.010	.014	.015	.015	.015	.016

Figure 3.3 Cumulative probabilities $D_c(k)$ in the monotonic case.



A similar procedure was followed in the free case where the model

$$\ln(1-D_c(k)) = \alpha'_c + (\beta' + \gamma' c)k \quad \text{(No. of categories } c=5,6,.,10)$$

for $D_c(k) > .75$ gave an excellent fit with a multiple correlation of .9999. The slope coefficient changed linearly in c and the estimates yielded $\hat{\beta}' = -.5773$ with a standard error of .0079 and $\hat{\gamma}' = .01906$ with a standard error of .0010. The estimates of $\hat{\alpha}'_c$ and their standard errors are given in table 3.9 and these results are summarized in the graph of figure 3.4.

Table 3.9. The regression coefficients α'_c in the free case for the model $\ln(1-D_c(k)) = \alpha'_c + (\beta' + \gamma' c)k$

$\hat{\beta}' = -.5773$ s.e.$(\hat{\beta}') = .0079$

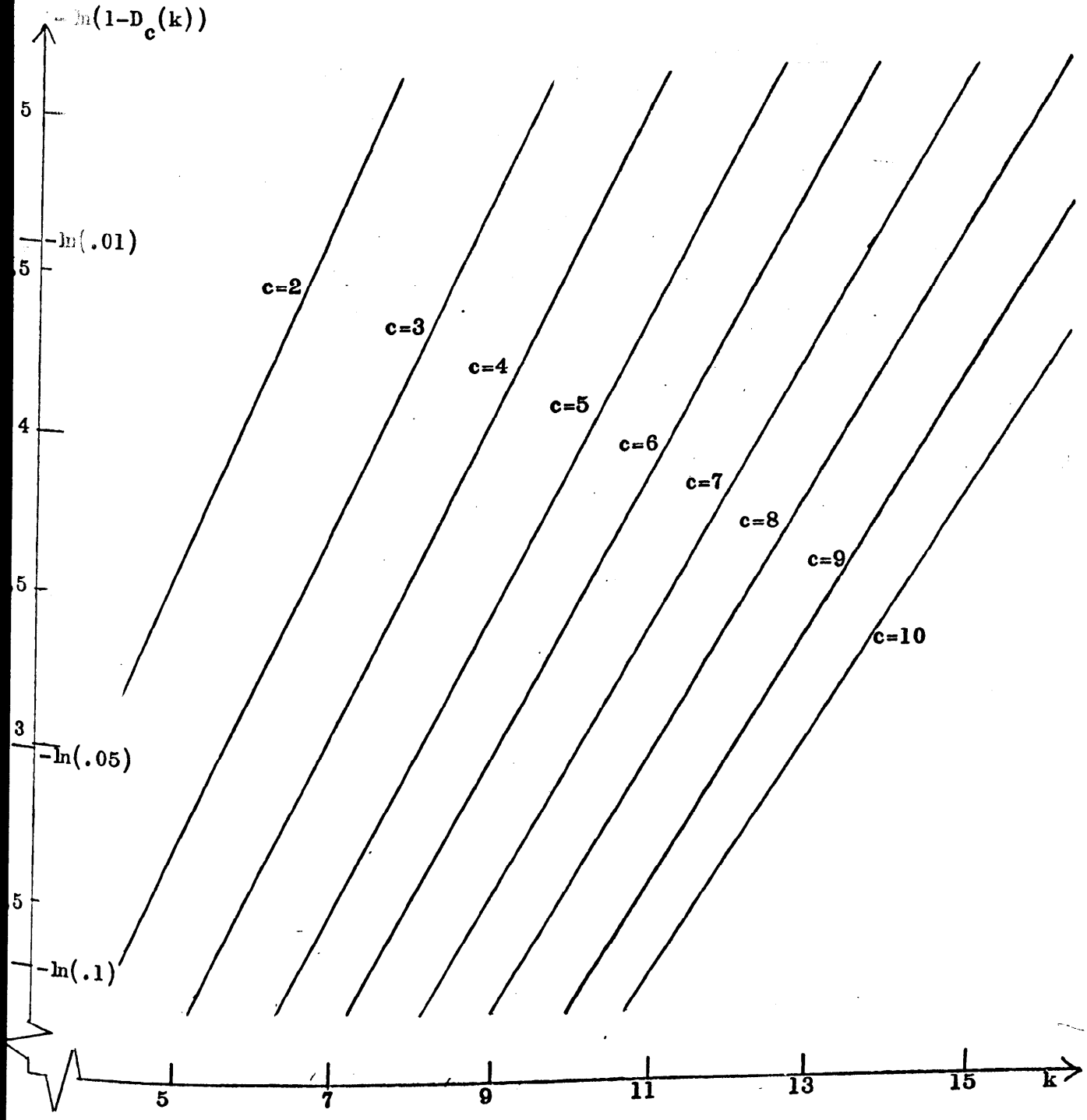
$\hat{\gamma}' = .01906$ s.e.$(\hat{\gamma}') = .0010$

No. of categories c	5	6	7	8	9	10
$\hat{\alpha}'_c$	.804	1.100	1.356	1.582	1.774	1.884
St. error of $\hat{\alpha}'_c$	.027	.021	.020	.022	.028	.036

Let it be immediately pointed out that these models are approximate since the relationship between k and $\ln(1-D_c(k))$ is not truly linear (if it were, the distribution would be χ^2_2 ie. exponential). In fact it is known a priori that these lines for different numbers of categories c cannot cross or touch each other except at $k = 0$.

/47...

Figure 3.4 Cumulative probabilities $D_c(k)$ in the free case.



These equations were used to give the critical values for the test statistic K for varying number of categories. They are presented and discussed in the next chapter.

/48...

This page intentionally left blank.

CHAPTER 4. Tables and Discussion of Results.

4.1 Cumulative distribution of K with equal allocation of observations to each category.

The right-hand tail of the distribution of K is described by the equations in section 3.2.4 when the number of categories c exceeds 4 and the observations are equally allocated to the available categories. These equations in k are solved to yield critical values of the 90%, 95% and 99% significance levels in tables 4.1 and 4.2.

Table 4.1. Critical values in the monotonic case with equal proportions of observations within each category.

	c								
Percentage point	2	3	4	5	6	7	8	9	10
90%	2.71	3.67	4.23	4.6	4.9	5.2	5.4	5.5	5.6
95%	3.84	4.89	5.50	5.9	6.2	6.5	6.7	6.8	6.9
99%	6.63	7.81	8.48	9.0	9.3	9.5	9.7	9.9	10.0

Table 4.2. Critical values in the free case with equal proportions of observations within each category.

	c								
Percentage point	2	3	4	5	6	7	8	9	10
90%	2.71	4.21	5.39	6.4	7.3	8.2	9.1	10.0	10.8
95%	3.84	5.49	6.76	7.9	8.8	9.8	10.8	11.8	12.6
99%	6.63	8.49	9.90	11.2	12.3	13.4	14.6	15.7	16.8

As expected the critical values increase as the number of categories increases. Although the rate of increase slows down, and especially in the monotonic case appears as if it might tend to some limit, this is not the case as shown in a similar problem by Hawkins (1975) who demonstrates that the expected value of this distribution follows the law of the iterated logarithm, and hence that the rate of increase is of the order $\log(\log c)$.

Note that a naive investigator who blithely applies a two-sample test to determine if there is a significant difference between his two new groups, would always be using the 2-category critical values in the above tables. As can be seen the loss in significance this would produce is mildly alarming in the monotonic case and very serious for free predictors.

The above results are for equal numbers of observations within each category. In the next subsection, 4.2, some alternative allocations of the observations will be considered and in 4.3 an upper bound to the critical values will be given that holds for any allocation of the observations to the various categories. As will be seen, the tables of this section seem to form "average" critical values over the different alternatives, and hence would offer good guidelines even for these cases.

/51...

4.2 Unequal allocation of observations to categories.

In the two category case, the asymptotic distribution of K is unaffected by differences in the numbers of observations within each of the two categories, as noted earlier in section 2.1.

There are very many ways of allocating observations to categories so two illustrative methods were chosen. One method was a symmetric allocation with progressively fewer observations in categories away from the centre, and the other a highly skewed allocation. The distributions resulting from such an allocation were evaluated by the quadrature program described in section 3.1 and are displayed in tables 4.3, 4.4 and 4.5 for 3, 4 and 5 monotonic categories respectively. The precision of the program was set to 10^{-4} for 3 and 4 categories, and 10^{-3} for 5 categories.

Table 4.3. Asymptotic cumulative distribution D(k) in three monotonic categories.

	Category			k							
	1	2	3	4	5	6	7	8	9	10	11
Equal	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	.9171	.9529	.9730	.9844	.9910	.9948	.9969	.9982
Symmetric	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	.9137	.9512	.9722	.9840	.9908	.9947	.9969	.9982
Skew	$\frac{8}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	.9224	.9557	.9745	.9852	.9914	.9950	.9970	.9983

/52...

Table 4.4. Asymptotic cumulative distribution D(k) in four monotonic categories.

	Category				k							
	1	2	3	4	4	5	6	7	8	9	10	11
Equal	1/4	1/4	1/4	1/4	.8862	.9343	.9619	.9778	.9871	.9924	.9956	.9974
Symmetric	1/8	3/8	3/8	1/8	.8763	.9291	.9592	.9766	.9864	.9921	.9953	.9973
Skew	7/10	1/10	1/10	1/10	.8970	.9403	.9653	.9797	.9881	.9930	.9959	.9976

Table 4.5. Asymptotic cumulative distribution D(k) in five monotonic categories.

	Category					k							
	1	2	3	4	5	4	5	6	7	8	9	10	11
Equal	1/5	1/5	1/5	1/5	1/5	.860	.918	.952	.972	.984	.990	.994	.997
Symmetric	1/16	1/4	3/8	1/4	1/16	.842	.908	.947	.970	.982	.989	.994	.996
Skew	6/10	1/10	1/10	1/10	1/10	.875	.927	.957	.975	.985	.991	.995	.997

The agreement is fairly good, especially as k increases. In all these cases the distribution when all categories have an equal number of observations lies between the distributions obtained when the observations are symmetrically and skewly allocated.

It is suggested that the special case of equal numbers of observations in each category is some sort of "average" result for the general case.

4.3 Upper bound to the critical values.

A simple method of getting a bound to the critical values is to use one of the Bonferroni inequalities, see for example Feller (1968, page 110). In particular, from the first inequality, the probability that one or more of the events $A_1, A_2, \ldots, A_n$ occur simultaneously cannot be greater than $p_1+p_2+\ldots+p_n$ where p_i is the probability that event A_i occurs. That is

$$P(A_1 \cup A_2 \cup \ldots \cup A_n) \leq p_1 + p_2 + \ldots + p_n \quad \ldots . \quad (4.1)$$

In our case, the probability that the K_i of section 2.1 (the test statistic for a given method of splitting the group) exceeds some critical value k_a is the right hand tail of the chi-square distribution with one degree of freedom $1-\chi^2(k_a)$.

/54...

Letting the event A_i correspond to $K_i \geq k_a$ it follows from 4.1 that

$$\begin{aligned} P(K \geq k_a) &= P(\max\{K_i\} \geq k_a) \\ &= P(K_1 \geq k_a \text{ or } K_2 \geq k_a \text{ or } \ldots \text{ or } K_n \geq k_a) \\ &= P(A_1 \cup A_2 \cup \ldots \cup A_n) \\ &= n(1 - \chi^2(k_a)) \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (4.2) \end{aligned}$$

where n is the number of permissible methods of splitting the group as given in section 2.1. Hence the Bonferroni upper bound to the $1-\alpha$ significance level is obtained from

$$\alpha \leq P(K \geq k_a) \leq n(1 - \chi^2(k_a))$$

ie. $\chi^2(k_a) \geq 1 - \alpha/n$

ie. $k_a \geq (\chi^2)^{-1}(1 - \alpha/n) \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (4.3)$

where $(\chi^2)^{-1}$ denotes the inverse transformation from the percentage point to the χ^2_1 deviate.

The upper limits to the 90%, 95% and 99% significance levels is given in tables 4.6 and 4.7 for the monotonic and free cases respectively. (See following pages).

Comparison of these tables with those of 4.1 and 4.2 show a remarkably close correspondence. The relative error decreases with increasing significance level, and increases with the number of categories. This is not too surprising from the form of the Bonferroni inequality. Note for example, that with ten categories at the 99% significance level the critical

/55...

Table 4.6. Bonferroni upper bound to the critical values for any allocation of the observations to categories in the monotonic case.

Number of Categories	Percentage Point 90%	95%	99%
2	2.706	3.841	6.635
3	3.841	5.024	7.879
4	4.529	5.731	8.615
5	5.024	6.239	9.141
6	5.412	6.635	9.550
7	5.731	6.960	9.885
8	6.002	7.237	10.169
9	6.239	7.477	10.415
10	6.447	7.689	10.633
11	6.635	7.879	10.828
12	6.805	8.052	11.004
13	6.960	8.210	11.165
14	7.104	8.355	11.314
15	7.237	8.490	11.452
16	7.361	8.615	11.580
17	7.477	8.733	11.700
18	7.586	8.844	11.813
19	7.689	8.948	11.919
20	7.787	9.047	12.020

values in tables 4.1 and 4.2 correspond to Bonferroni significance levels of 98.6% and 97.9% so that the Bonferroni estimates at this level are not overly conservative.

/56...

Table 4.7. Bonferroni upper bound to the critical values for any allocation of the observations to categories in the free case.

Number of Categories	Percentage Point 90%	95%	99%
2	2.706	3.841	6.635
3	4.529	5.731	8.615
4	6.002	7.237	10.169
5	7.361	8.615	11.580
6	8.675	9.945	12.934
7	9.974	11.256	14.266
8	11.271	12.562	15.588
9	12.569	13.868	16.909
10	13.872	15.178	18.231
11	15.180	16.491	19.555
12	16.492	17.809	20.882
13	17.809	19.130	22.212
14	19.130	20.455	23.544
15	20.455	21.784	24.879
16	21.784	23.115	26.217
17	23.115	24.449	27.556
18	24.449	25.786	28.898
19	25.786	27.125	30.241
20	27.125	28.466	31.587

Since the Bonferroni upper limits holds true for any allocation of the observations to the categories, and because of the ease of their calculation it seems plausible that these values should be used in significance testing in those cases where the exact critical values have not been tabulated - until

/57...

perhaps such time as a fast approximation of sufficient accuracy can provide more exact figures. The argument for using the Bonferroni upper bounds is strengthened by the apparent closeness these figures bear to those of the special cases that we have tabulated. In anticipation of such use, the tables of 4.6 and 4.7 are extended up to twenty categories.

4.4 Evidence of applicability to non-infinite samples.

Strictly, the asymptotic results are only applicable to infinite samples. As mentioned in the introduction to this thesis, A.I.D. should be performed on large data bases, and subgroups should not be analysed when containing very few observations.

With only two categories and a dichotomous dependent variable, the problem reduces to using chi-square with 2x2 contingency tables. This approximation is acceptable for sample sizes greater than forty, see for example Grizzle (1967).

With a three category predictor and equal numbers of observations in the first and third categories the exact permutation distribution of K can be obtained in the monotonic case for a dichotomous dependent variable (taking values -1 and 1 say) with equal numbers of observations in each part of the dichotomy. For a total of N observations and N_i observations in category i, this case has $N_1=N_3$. Consider a random walk from (0,0) to (N,0) in the notation of Feller (1968,

/58...

Chapter 3), where (t,x) denotes being at position x at epoch t. At each observation, a step is taken one unit up if the dependent variable is 1 and a step is taken down if it is -1. The position at epoch N_1 determines the number of -1's and 1's in the first category, so if the position (N_1,x_1) is reached, the first category had x_1 more 1's than -1's (note that x_1 may be negative) and the mean within this category is

$$\bar{Y}_{11} = (\tfrac{1}{2}(N+x_1) - \tfrac{1}{2}(N-x_1))/N_1$$
$$= x_1/N_1 \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (4.4)$$

If the position at epoch N_1+N_2 (ie. N_3 from the end) is x_2 then by symmetry the mean of the third category is

$$\bar{Y}_{22} = x_2/N_3 = x_2/N_1 \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (4.5)$$

since $N_1=N_3$ in this case. Also because of the choice of the dependent variable, the grand mean $\bar{X}=0$ and the total variance $S^2=1$, so that the test statistic obtained by splitting after the first category is from 2.2

$$K_1 = M_1 N(\bar{Y}_{11}-\bar{X})^2/M_2S^2 = N_1Nx_1^2/N_1^2(N-N_1)$$
$$= Nx_1^2/N_1(N-N_1) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (4.6)$$

and by symmetry, the test statistic obtained by splitting after the second category is

$$K_2 = Nx_2^2/N_3(N-N_3)$$
$$= Nx_2^2/N_1(N-N_1) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (4.7)$$

/59...

The test statistic K is the maximum of K_1 and K_2 and hence is determined by the absolute maximum of x_1 and x_2, that is the maximum deviation from the origin at epochs N_1 and N_1+N_2. The paths of interest from $(0,0)$ to $(N,0)$ are obtained by selecting N_1 observations out of the N for the first segment, and then N_2 for the second yielding N_3 for the third, and the total number of such paths are given by the multinomial coefficient $N!/(N_1!N_2!N_3!)$.

If $K=0$ then $K_1=K_2=0$ which occurs only if $x_1=x_2=0$. These paths are obtained by selecting $\frac{1}{2}N_1$ of the $+1$'s and $\frac{1}{2}N_1$ of the -1's for the first segment, $\frac{1}{2}N_3$ of the $+1$'s and $\frac{1}{2}N_3$ of the -1's for the third segment so that

$$\text{Prob}(K=0) = \binom{\frac{1}{2}N}{\frac{1}{2}N_1}^2 \binom{\frac{1}{2}(N-N_1)}{\frac{1}{2}N_3}^2 \Big/ (N!/(N_1!N_2!N_3!)) \quad \ldots\ldots\ldots\ldots (4.8)$$

if N_1 is even. When N_1 is odd it is impossible to return to the origin in N_1 steps and we may follow Feller's convention of defining the binomial coefficient

$$\binom{n}{r} = 0 \quad \text{if } r \text{ is not a non-negative integer} \quad \ldots\ldots\ldots\ldots (4.9)$$

When K is not zero, then either x_1 or x_2 is not zero, and because of the symmetry only the case $x_1^2 \geqslant x_2^2$ need be considered, the final probabilities being doubled in compensation when $x_1^2 \neq x_2^2$. Let the absolute maximum attained at epoch N_1 be $x=x_1$ so that the first segment contains $\frac{1}{2}(N_1+x)$ of the $+1$'s and $\frac{1}{2}(N_1-x)$ of the -1's which can

/60...

be chosen in

$$\binom{\frac{1}{2}N}{\frac{1}{2}(N_1+x)}\binom{\frac{1}{2}N}{\frac{1}{2}(N_1-x)} \quad \text{ways,}$$

leaving $\frac{1}{2}(N-N_1-x)$ of the $+1$'s and $\frac{1}{2}(N-N_1+x)$ of the -1's.

If $x_2=0$ then the third category must have $\frac{1}{2}N_3$ of the $+1$'s and a similar number of -1's which can be selected in

$$\binom{\frac{1}{2}(N-N_1-x)}{\frac{1}{2}N_3}\binom{\frac{1}{2}(N-N_1+x)}{\frac{1}{2}N_3} \quad \text{ways.}$$

When $x_2\neq 0$ then we need only consider the case $x_2>0$, and multiply the corresponding probability by two because of the symmetry of $x_1\gtrless 0$ $x_2<0$ and $x_1\lessgtr 0$ $x_2>0$. The third category must then contain $\frac{1}{2}(N_3+x_2)$ of the $+1$'s and $\frac{1}{2}(N_3-x_2)$ of the -1's that can be selected in

$$\binom{\frac{1}{2}(N-N_1-x)}{\frac{1}{2}(N_3+x_2)}\binom{\frac{1}{2}(N-N_1+x)}{\frac{1}{2}(N_3-x_2)} \quad \text{ways}$$

and all cases will be considered if we sum over all positive x_2 that are not greater than $|x_1|$.

Summarizing these results and including 4.8 gives

$$P(K = Nx^2/(N_i(N-N_i)))$$

$$= \binom{\frac{1}{2}N}{\frac{1}{2}N_1}^2\binom{\frac{1}{2}(N-N_1)}{\frac{1}{2}N_3}^2 \Big/ (N!/(N_1!N_2!N_3!)) \quad \text{if } x=0$$

$$= \frac{2\binom{\frac{1}{2}N}{\frac{1}{2}(N_1+x)}\binom{\frac{1}{2}N}{\frac{1}{2}(N_1-x)}\sum_{x_2=0}^{x}\binom{\frac{1}{2}(N-N_1-x)}{\frac{1}{2}(N_3+x_2)}\binom{\frac{1}{2}(N-N_1+x)}{\frac{1}{2}(N_3-x_2)}\delta(x_2)}{(N!/(N_1!N_2!N_3!))}$$

$$\text{if } x=\pm 1,\pm 2,\ldots,\pm N_1$$

$$= 0 \quad \text{otherwise} \qquad \ldots\ldots\ldots\ldots\ldots\ldots \quad (4.10)$$

where $\delta(x_2) = 1$ if $x_2 = 0$ or x; $= 2$ otherwise,

and Feller's convention 4.9 is followed.

/61...

The function 4.10 is plotted in figure 4.1 (see next page) for various sample sizes N (taking $N_1=N_2=N_3$), where the asymptotic distribution is also drawn for comparison. It can be seen that in the tail area of the distribution the asymptotic result gives a good approximation even for small sample sizes.

4.5 Extension to several predictors.

If all the predictors are statistically independent then there is no difficulty in extending the results. If there are p predictors and critical values are chosen so that α_i is the probability under the null hypothesis that the i^{th} predictor is significant, then the overall significance or the probability that one or more of the predictors will exceed its critical value is

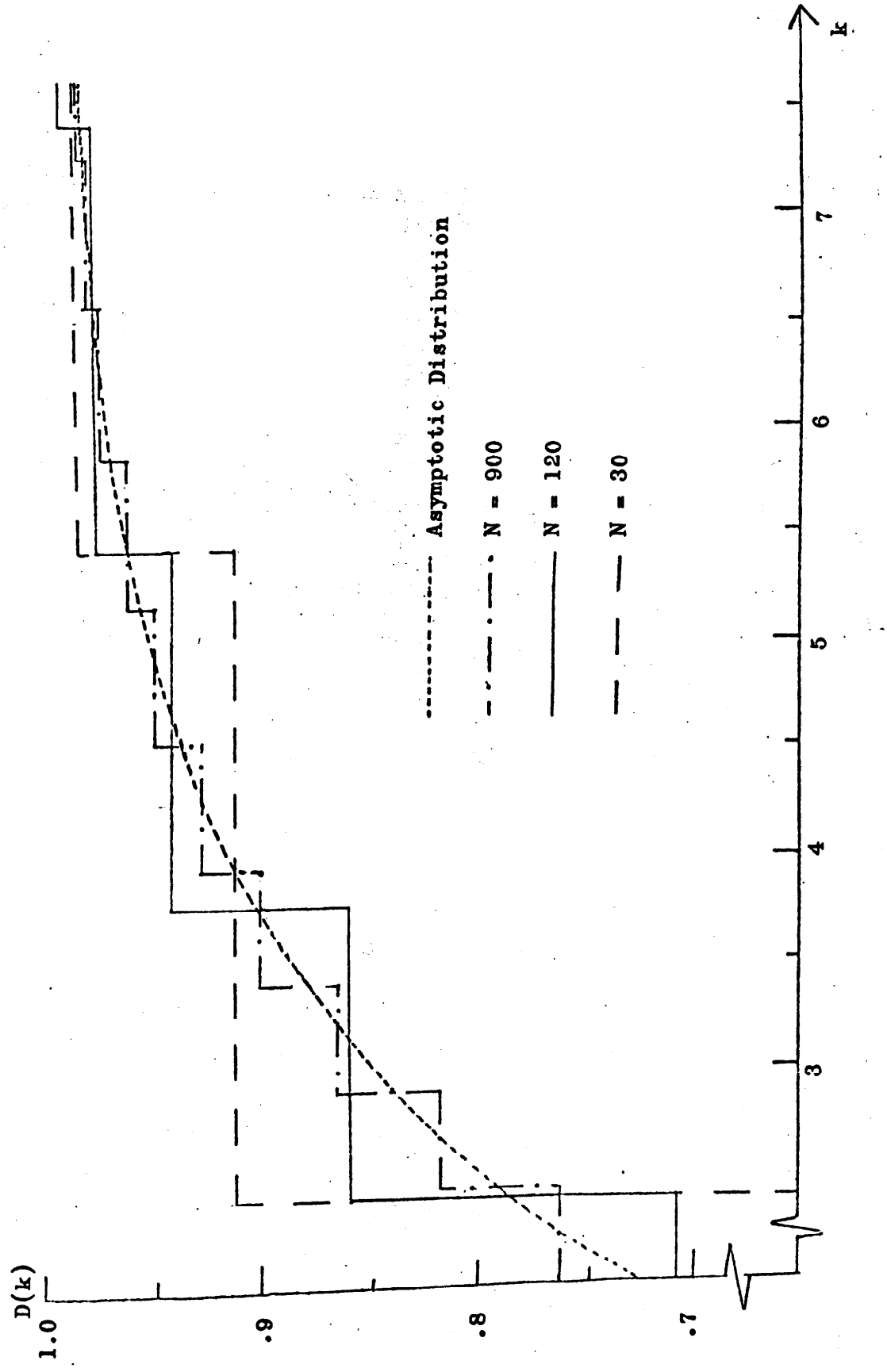
$$1 - \prod_{i=1}^{p}(1-\alpha_i)$$

by standard theory, see for example Miller (1966, page 8).

When two or more predictors are not statistically independent there are many serious difficulties, not the least being a definition of dependence between predictors when one or more of them are not monotonic. While the assumption of independence may be used as above to yield a "very approximate" overall significance level, the Bonferroni inequality 4.1 can again be invoked to give a lower bound. This conservative approach yields the overall significance level,

/62...

Figure 4.1 Cumulative null distribution of the test statistic K for three categories with equal numbers of observations within each category, for various sample sizes.



that is the probability that one or more of the predictors will exceed its critical value, as being not greater than

$$\sum_{i=1}^{p} \alpha_i$$

Again, the bound to the overall significance level may be pre-set and the α_i chosen accordingly.

If the predictors are 'positively' related in that the significance of one predictor implies that other predictors will also be significant, this Bonferroni bound to the global significance level is highly conservative. In the extreme case where all the predictors are either all significant or all not significant the true overall size of the test is $\max\{\alpha_i\}$ and in the special case of identical α_i each equal to α, then the size of the test is α while the Bonferroni bound is $p\alpha$. In some practical applications predictors may be similar because of inherent relationships or since they may be different measures of the same quantity - for example in a sociological setting 'Date of marriage' and 'Number of children' are both related, and to a certain extent could also be crude measures of 'Age'.

However, if the predictors are not 'positively' related then the Bonferroni bound improves as an estimate of the global significance, and in particular if the predictors are independent the Bonferroni bound is by algebraic expansion a first approximation to the significance level since

$$\prod_{i=1}^{p} (1-\alpha_i) \approx 1 - \sum_{i=1}^{p} \alpha_i$$

/64...

for α_i small, the neglected terms being of order $\alpha_i \alpha_j$.

In any event, by fixing the overall size of the test and using the Bonferroni bound, if there are many predictors the α_i become comparatively small, and if they are less than .01 say, then the Bonferroni upper bounds of the individual critical values become more exact as mentioned towards the end of section 4.3. Thus decreasing the bound at the final stage tends to reduce the conservativeness of the estimates at the prior stage.

/65...

CHAPTER 5. Summary and Conclusion.

5.1 Summary of results.

The figures presented in this thesis offer a guide to whether results obtained by an A.I.D. analysis are in some way significant or not. As is usual with tests of significance there is really no absolutely determined level beyond which results are "significant" and within which results are "attributable to chance". After all, A.I.D. is predominantly a search procedure, and informed commonsense should prevail in the evaluation of the results. The presentation of the Bonferroni upper bounds provides a region of fairly secure confidence where results are established as being significant. Crudely the two-sample tests (corresponding to the two category case), and better, the asymptotic results for a single predictor in the case of equal allocation to each category, provide critical values below which predictors are almost certainly insignificant. This leaves a grey-area within which no categorical conclusion may be reached - an area which is usually left open to further investigation and research.

The theory has been based on the permutation distribution of the test statistic, and hence is applicable to any distribution satisfying the conditions given in Puri and Sen (1971, page 70ff) who review and extend the literature on permutational limit

/66...

theorems. In particular the content of this part of the thesis is applicable to a distribution-free analogue of A.I.D. where the dependent variable is transformed to a set of ranks or normal scores. It is expected that the asymptotic results will hold more closely for small samples of transformed data than if the dependent variable is left untransformed when it is highly skewed or otherwise badly behaved. In particular if random normal scores are used, see for example Bradley (1968), then the group's theoretical variance is known and need not be estimated by S^2, and the asymptotic results become exact for all sample sizes. Note that if it is decided to transform the data, then the data should be retransformed within each new group formed by a split, since it is unlikely that the new groups will retain the same distribution as the parent group - in fact in the ideal case where the split divides the dependent variable into an all high group and an all low group only the rectangular distribution will retain its shape after the split.

5.2 Conclusions.

It is hoped that the availability of a valid measure of the significance of A.I.D. results will increase the confidence of those who use A.I.D. and will allay some of the many fears of those who oppose it as expressed by Professor G.A.Barnard in the discussion on a paper by Stone (1974): "... but nowadays with more and more apparently sophisticated computer

/67...

programs for social science, failure to take account of possible sampling fluctuations is leading to a glut of unsound analyses which already seems to be bringing statistical method in that area into unjustified disrepute. I have in mind procedures such as AID, the automatic interaction detector, which guarantees to get significance out of any data whatsoever." It is hoped that this part of the thesis goes some way in providing the necessary validation.

In view of these results the A.I.D. algorithm could be modified to provide a more parsimonious description of the data, by splitting a group on the most significant predictor rather than on that predictor which explains the most variance.

Ideally it would be helpful to be able to test the complete analysis for significance. A conservative test could be based on the Bonferroni inequality in a similar manner to the way it has been used in other stages of the analysis. Otherwise, a step by step procedure could test each group's split independently of the other's. Two possibilities exist: either insignificant splits are immediately discarded as soon as they are found, or the complete A.I.D. analysis conducted first and then the smallest groups recombined if necessary. In the first case it is reasonable to split on the most significant predictor, while in the second case there is the advantage that if an insignificant split gives rise later to a significant one this will not be lost. These problems are

/68...

not peculiar to A.I.D. but are paralleled by similar problems in stepwise regression where, as far as this author is aware, the problem of significance testing at each step, at what level to accept a new predictor or reject an old one, has not yet been fully resolved.

While A.I.D. is extremely useful it has certain limitations. Various extensions have been made to A.I.D., for example the case of multiple dependent variables has been investigated recently by Gillo and Shelly (1974). Two limitations which we have found irksome in practice are firstly that the dependent variable must be ordinal where nominal is desirable, and secondly that the splits are binary where a three or more way split is preferable. The latter problem is not prima facie too serious since A.I.D. could initially perform one of the splits and subsequently may re-use the predictor and bring the other split points into play. However from the user's viewpoint it is preferable to be aware of the trichotomous nature of the predictor as soon as it is considered rather than be made aware of its form only at a later stage. An attempt has been made to solve the first problem of a nominal dependent variable by Messenger and Mandell (1972) who consider an A.I.D. type method that is unfortunately not compatible with standard A.I.D. in the case of a dichotomous dependent variable.
Part II of this thesis proposes a technique for removing these two limitations and contains some potentially useful by-products.

PART II

CHAPTER 6. Introduction and Orientation.

The second part of this thesis introduces a method of data analysis, similar in structure to A.I.D., for analysing completely categorized data - ie. the dependent variable is purely nominal ("free" in previous terminology, though this could be extended to a "monotonic" or an otherwise restricted dependent variable). Basically, A.I.D. uses the two-sample t-test (equivalently the between-sum-of-squares) as its criterion, which is logical since the aim is to maximize the variance between the two groups formed by a split. Examining a nominal dependent variable according to a categorized predictor immediately brings to mind the chi-square test of association in a contingency table, and indeed this is the criterion used in this technique.

A number of other extensions to conventional A.I.D. are also included. Instead of confining a single stage of the analysis or the investigation of the effect of a particular predictor to a simple two-way split, criteria are presented to allow multi-way divisions. In the case of the analysis with which this part is concerned, the chi-square test can be used routinely to examine multi-way splits as will become clear in the sequel. The analogue in conventional A.I.D. would be to use analysis of variance (ANOVA) to examine a k-way split. The multi-way split raises two points. Firstly, the method does not produce

/70...

a different subgroup for each category of the predictor, but combines those categories of the dependent variable which have similar distributions, leading to a more parsimonious description of the data than would otherwise be the case. Secondly, the method may show (and often does in practice) that a predictor is indivisible according to the criteria. This is felt by the author to go a long way towards reducing the possibility of abuse of the method by inexperienced researchers and to allaying many fears expressed about the validity of A.I.D. type techniques.

The concepts of free and monotonic predictor variables are retained in this method and a new type of predictor is introduced. It has often been found that an otherwise monotonic predictor has a category that cannot be placed logically a priori on the monotonic scale. This situation arises naturally with missing information in survey data and "don't know" or "undecided" in behavioural applications. Such a predictor is defined as monotonic with a single "floating" category that may appear alone or combined with any other group as found appropriate in the analysis. A generalization of this problem is also considered. The new types of predictors are not peculiar to the new method and could equally well be implemented in conventional A.I.D., although this thesis only considers the implications arising in the case of a nominal dependent variable.

/71...

Finding the optimum w-way split of a c-category predictor is not trivial, but can be performed routinely by dynamic programming in the monotonic case, and with slight modifications for a predictor with a "floating" category. Analysing a large number of predictors over various subsets as occurs in A.I.D. type analysis renders this approach too tedious for general use and hence a stepwise procedure for analysing each predictor was evolved. The method is consistent in finding well-defined divisions and experience indicates that its results are very similar to the optimum solution.

If in the case of a dichotomous dependent variable, the conventional A.I.D. analysis and this method are compared, when a binary split of a predictor occurs the split is usually defined in the same way and results in the same test statistic value. A continuous dependent variable cannot be analysed directly by this new method, but could be after a suitable grouping of the dependent variable. While this may at first sight appear wasteful, it would allow the extensions to A.I.D. provided by this analysis to be used (until such time as conventional A.I.D. is updated), and it may also provide a convenient solution to a highly skewed or otherwise badly behaved dependent variable.
In a different context, Hill (1974) has found practical examples where categorized data perform better in an analysis than the original quantities.

Finally we note that, despite the safeguards provided by this system of data analysis, it still remains a large sample technique - it could well prove overwhelmingly conservative for small data sets.

/72...

6.2 Previous work in this area.

The problem of analysing a categorized dependent variable has received attention in the literature. Two main references approach the problem in the same direction as this thesis - via a heuristic A.I.D. type approach.

The first is Press et al. (1969) who describe an on-line data analysis system called IDEA (Inductive Data Exploration and Analysis) that produces a decision tree similar to an A.I.D. dendrogram. As its name implies IDEA is a computer on-line technique, in which the researcher decides how the analysis should proceed from information provided by the program. Press et al. compare their technique with the standard A.I.D. of Morgan and Sonquist (1963) and note three major differences. As has been described earlier A.I.D. only performs binary splits of the data, IDEA "employs routines which allow relationships in the data to determine the number of branches at each node." A.I.D. assumes that the dependent variable is measured on an interval scale while IDEA does not make this restriction. Finally A.I.D. works completely automatically in that it decides for itself the best split, performs the split and continues the analysis; IDEA on the other hand permits the researcher to examine the output and decide for himself how he would like the program to proceed. Subsequent to the publication of the article by Press et al., there appeared a new version of A.I.D. called AID-III in Sonquist et al. (1971) which allows the user to direct the analysis

/73...

as in IDEA, but which is cumbersome since AID-III is batch-processing orientated rather than interactive.

The improvement of IDEA over A.I.D. lies mainly in its ability to handle non-interval scale dependent variables and in the information that can be provided by multi-way splits. However, no effort appears to have been made to assess the validity of the results and the possibility of a user directing the program how to proceed does nothing to allay the fears of those that consider the output of such routines highly suspect.

Press et al. also compare their routine with that of Hunt et al. (1966) who have written a Concept Learning System (CLS). The similarity to IDEA and A.I.D. seems extremely tenuous and perhaps based on the common use of decision trees. The contribution to our particular problem is negligible since Hunt et al. have focussed their attention on perfectly logical data (sentences) that do not permit contradictions. In the terminology of data analysis this means that a data vector with identical predictors must necessarily have identical dependent variables - clearly an impractical requirement for the data that usually occur in practice. Further, their criteria are not mathematically based and though they briefly consider the "chi-square criterion" and the "maximum transmitted information" neither was programmed in the studies they report, and they are dismissed briefly by (page 95):

/74...

"These are more elegant mathematical and statistical criteria on which subproblem selection may be based. They have the advantage that they can be used in other than binary branching trees, and the disadvantage that they require more computation."

The other main reference is to a computer program called Theta A.I.D., abbreviated to THAID. A preliminary discussion is given in Messenger and Mandell (1972) and a more complete description in the monograph of Morgan and Messenger (1973). This is an A.I.D. type technique specially designed for a nominal dependent variable. As in conventional A.I.D. only binary splits are considered, so that multi-way splits can only be revealed at a later stage of the analysis if a predictor is re-used. The criterion for comparing various splits is based on the "theta statistic" that is defined as "the proportion of the sample classed correctly when using the optimal-prediction-to-the-mode strategy" (Morgan and Messenger, 1973, page 12). That is, for each possible split the sum of the number of observations in each modal category is maximized.

In the case of a dichotomous dependent variable where A.I.D. and THAID are both applicable the theta criterion will yield in general a different solution to conventional A.I.D. Morgan and Messenger consider this an advantage mainly on intuitive grounds since they note (page 22): "Roughly speaking, AID tends to split giving outlying subgroups of small size much more than does THAID." They performed a computer simulation

/75...

that compared various criteria for splitting and showed that the theta statistic did not correlate highly with any of their other criteria, and in particular was not closely related to the chi-square criterion which we will show later to be identical to the standard A.I.D. criterion for a dichotomous dependent variable.

We disagree with Morgan and Messenger and feel that an extension of A.I.D. to a nominal dependent variable should as a matter of course be compatible with standard A.I.D. in the case of a dichotomous dependent variable where the two methods overlap. That is, the ultimate objective of partitioning the data into maximally internally homogeneous groups should yield identical groups in the case of a binary split. The basis for their observation that A.I.D. tends to produce highly unbalanced splits is not clear. Under the alternative hypothesis that a group should be split into two subgroups, any consistent method should be able to do this and find the correct split point even if this means producing an unbalanced split. It therefore seems reasonable to assume that the phenomenon they have observed arose in situations where the null hypothesis held, that in fact the group is homogeneous and should not be split. In this case they would be merely observing the result of Hawkins (1975) who showed that under the null hypothesis the position of the split point followed a U-shaped distribution which tended to the arc-sine distribution as the number of categories in the predictor increased.

/76...

Hence it is not at all surprising that the A.I.D. procedure should produce unbalanced splits. This is really no indictment of the A.I.D. criterion since under the null hypothesis no split should be produced and hence there is no reason for preferring a balanced split to an unbalanced one.

Another reason for not preferring the theta criterion to the standard A.I.D. criterion (which is equivalent to the chi-square criterion) is the lack of knowledge of its theoretical behaviour. A guide to the user has been provided by Morgan and Messenger (pages 20, 21) who replicated ten times a Monte Carlo experiment on a dichotomous dependent variable with eight 5-category predictors with a uniform allocation of observations. The change (with an upper and lower bound) in the theta criterion against the number of split groups is shown on a graph. They claim that "the choice of distribution type, nominal predictors, number of codes and number of predictors was made to make the results conservative and hence an upper bound on the growth of $\Delta\theta$." Their claim of conservativeness is obviously dependent on the number and characteristics of the predictors which one is accustomed to use. The only extension to their results that they offer is that further simulation has indicated that their figures are inversely proportional to the square-root of the group size.

As will be shown subsequently, the theory in part I of this thesis is applicable to dichotomous dependent variables

/77...

when using the chi-square criterion, and the similarity of the Bonferroni approximation to the exact critical values leads to the incorporation of significance testing as part and parcel of the program, not only to decide the number of subgroups into which a group should be partitioned but also whether a group should be split at all. Another consequence of using significance testing in the decision making process will be to counteract the undue weight afforded to predictors with a large number of categories that can usually "explain" more than predictors with fewer categories because they allow more possibilities for splitting and hence capitalize on chance to a far greater extent.

6.3 Summary of contents of part II.

After a discussion of the general aims of this analytic technique, the computer implementation will be presented. The various types of predictors will be considered: the monotonic and free predictors that have been encountered earlier, together with a floating predictor which often occurs in practical experience. A number of hypothetical problems concerning the analysis will be answered by appropriate numerical examples.

A short chapter than follows presenting the results of a Monte Carlo simulation under the null hypothesis with special reference to each type of predictor and varying numbers of categories

/78...

for the dependent variable and the predictor variable.

The penultimate chapter gives a sample analysis of some "live" data that illustrate the various options within the computer program.

The final chapter summarizes the main conclusions and considers various extensions for future research. Appendix B gives an annotated listing of the computer programs together with instructions for their use. Appendix C gives a reprint of the computer output for the analysis of the example mentioned above.

/79...

CHAPTER 7. The Method of Analysis.

7.1 Outline of the technique.

Most authors provide their computer programs with acronyms both to provide a simple reference to the program and to describe the main features of the technique in a simple mnemonic. If the practice is followed of Morgan and Messenger (1973) who created THAID, and of Gillo (1972) who created MAID, then a suitable name for the technique to be presented is CHAID - chi-square A.I.D. As this name would imply, the criterion on which the analysis is based is the chi-square statistic. This is a natural measure since the nominal dependent variable and a categorized predictor allow the data to be summarized in a contingency table as in figure 7.1.

Figure 7.1. Summary of data for one predictor.

		Dependent variable					
		1	2	3	..	..	d
Predictor	1	0_{11}	0_{12}	0_{13}	..	..	0_{1d}
	2	0_{21}	0_{22}	0_{23}	..	..	0_{2d}
	3	0_{31}	0_{32}	0_{33}	..	..	0_{3d}
	.	..	..	..	..	..	..
	.	..	..	..	..	..	..
	c	0_{c1}	0_{c2}	0_{c3}	..	..	0_{cd}

0_{ij} denotes the number of observations that are classified into category i of the predictor and have characteristic j of the dependent variable.

/80...

In practice there is more than one predictor and hence more than one 2-way contingency table is created. As in part I of this thesis, the preliminary discussion will be concerned with a single predictor, the problem of multiple predictors being postponed.

As in conventional A.I.D., CHAID proceeds in a stepwise fashion - first finding the most significant predictor of the total group, then splitting the total group according to this predictor, and then analysing each of these subgroups in turn and where possible further splitting these subgroups and re-analysing these smaller subdivisions.

At each stage of the analysis the best way of subdividing the data according to each predictor is ascertained. That is, the categories of the predictor are combined, due cognizance being taken of the type of the predictor as this determines the permissible groupings of its categories, so as to produce the contingency table with the highest significance level according to the chi-square test, see for example Conover (1971). Note that this implies that there are a sufficient number of observations to ensure the validity of this test. If this is not the case then some other criterion could be used, such as Fisher's exact test (Conover, 1971, page 163ff), but we will not pursue this possibility.

/81...

In principle the optimum method of grouping the categories of a predictor can be found by dynamic programming, but while it may be computationally feasible on a computer for a single predictor and a single stage of the analysis, it would be unrealistic in a practical example where there is more than one predictor and the analysis extends over a number of stages. This problem is similar to that of multiple regression where a search is made for the "best" subset of predictors. Here, the solution generally accepted is that produced by "stepwise" regression, see for example Efroymson (1965). A similar stepwise procedure was applied to our problem and found to work quite satisfactorily. Thus the analysis of a predictor was further broken down into stages as follows.

Firstly, the "full" contingency table with c rows (for a c-category predictor) is calculated. Pairs of rows (as determined by the type of the predictor) are examined and the most insignificantly different pairs of rows are merged if their significance does not exceed some predetermined amount. That is, the various $2 \times d$ (for a d-category dependent variable) subtables are examined and the least significant subtable "collapsed" unless its significance reacnes some pre-assigned value.

The $(c-1) \times d$ table is similarly analysed to determine if two further rows can be merged. When the merging criterion is not met, or when all rows have been merged, all compound rows are re-examined to determine if they can be split up again. That is, a compound row is examined to see if a

/82...

2 x d table that can be formed from it, is significant at some pre-assigned value. This process is continued until no two rows are sufficiently similar to be merged, and no compound row can be decomposed into two significantly different rows. In order that such a stable solution should be reached it is necessary to ensure that the criterion for splitting a compound row is stricter than the criterion for merging two rows. That is, the critical value for a split must have significance greater than that required to prevent a merge. (To avoid possible program "looping", the routines will not attempt to split a compound row unless a stricter criterion than the merging criterion is provided).

The logic of the above method is displayed in the flowchart of figure 7.2. (See next page).

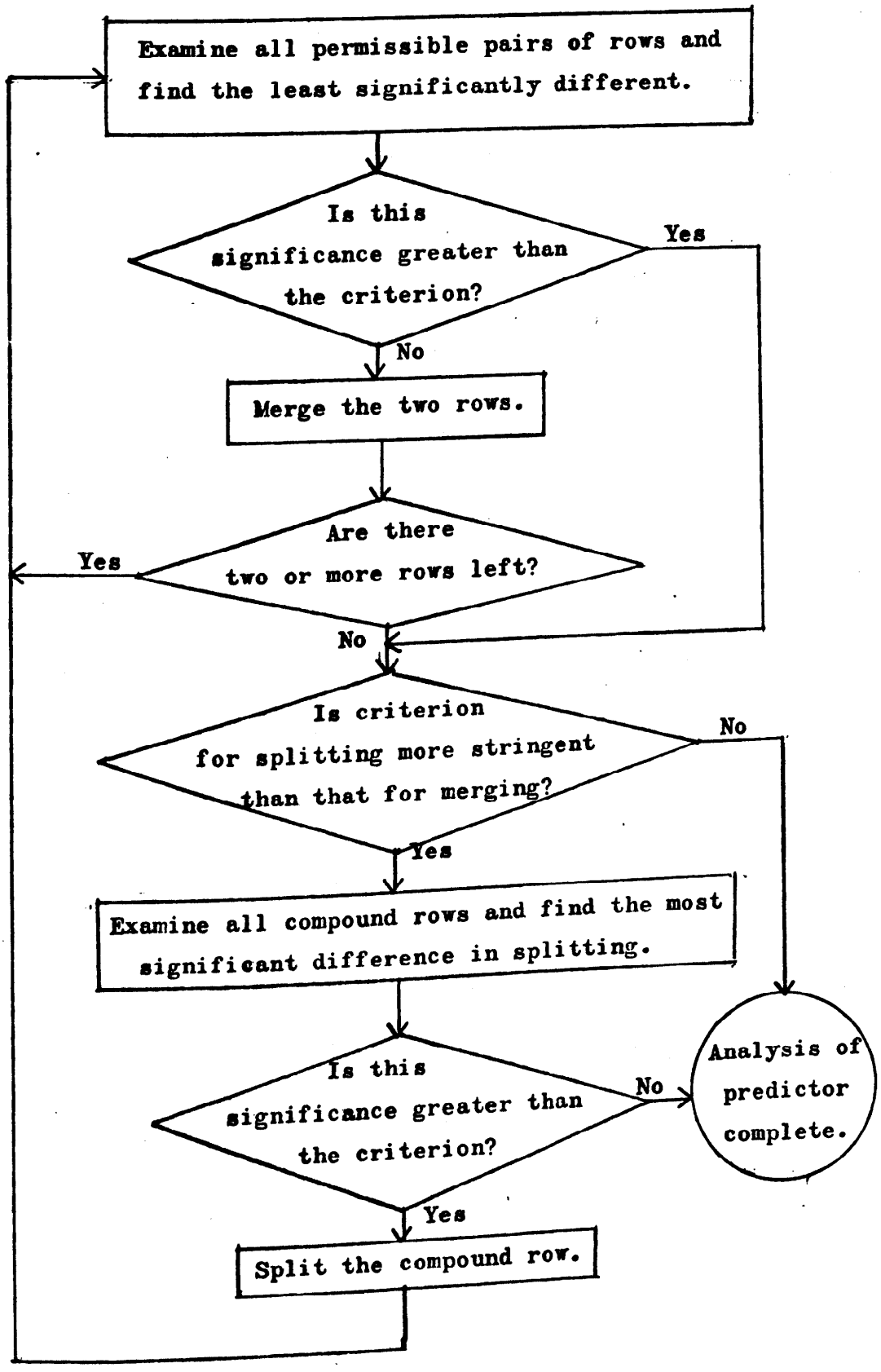
In the asymptotic case the procedure must converge, as will be proven in 7.1.3 below, where the implications for finite samples will also be considered.

An added advantage of this stepwise approach in comparison with the global "best splits" direct method is that the merging operation can be continued until there are no rows left to merge - an indication that none of the rows is significantly different from the rest.

It may be wondered whether there is any gain from splitting a compound row. The nature of the algorithm precludes a split that would "undo" a merge, that is, recreate a situation

/83...

Figure 7.2. Flowchart depicting logic of analysis for a single predictor.



that existed before a merge took place. In practice it has been found that splitting a compound row is a comparatively rare event, but must be considered if this approach is to give near optimum results. Situations which result in a compound row being split will be illustrated in section 7.2 where the various types of predictors will be discussed.

The procedure of merging categories is not new, but has been used in other contexts. The reverse procedure of starting with all categories combined and then splitting this compound group has also received attention. Anderson (1971, page 42 et passim) considers both these approaches and demonstrates a disadvantage of the latter procedure - that the splitting process may in certain circumstances terminate too soon and thus produce too few groups. McGee and Carleton (1970) while dealing with the problem of piecewise regression also start at the atomic level and "merge" until all observations are combined in a single group.

Having obtained the best "reduced" contingency table that the method can find, the problem arises of testing it for significance. The solution is straightforward if the predictor has only two categories since the chi-square test for contingency tables is applicable immediately. Similarly, if no reduction of the original contingency table has taken place, a chi-square test with the appropriate degrees of freedom can again be used. The problems really arise when "reduction" of the original

/85...

table has taken place, since the resultant table is the "best" possible for its size. Since the case of a dichotomous dependent variable is equivalent to standard A.I.D. if the table is reduced to two rows (to be demonstrated later), the results of part I of this thesis may be applied. Noting the similarity between the Bonferroni critical values and those obtained by the exact theory it was felt that the Bonferroni results could be used as a conservative first approximation to the exact results even in those cases where the exact results are unknown. This assumption in fact has been partially validated by the computer simulation presented in chapter 8. In effect, the procedure is to determine the number of "reduced" tables of the same size that could be derived from the original table, a quantity that depends on the type of the predictor, and use it in the Bonferroni inequality to obtain a bound for the multi-significance level. The formulae for calculating these multipliers are derived in the next section dealing with the various types of predictors.

Once the most significant partition has been found, then given that it meets some preset criterion, this partition is actually performed on the data. As in conventional A.I.D., each of these divisions is re-analysed, further partitions being formed where appropriate. The sequence of analysis and partitioning is continued until either no more significant divisions can be found, or some other stopping criterion is reached.

/86...

Some properties and an alternative to the method that has been outlined above are considered in the next four sub-sections.

7.1.1 Equivalence with A.I.D. in the case of a dichotomous dependent variable.

Let the "reduced" 2 x 2 table be as in figure 7.3, where the first and second rows of the predictor are possible compound rows corresponding to the two parts of the split.

Figure 7.3. Dichotomous dependent variable.

		Dependent variable 1	2	Total
Predictor	Row 1	a	b	a + b
	Row 2	c	d	c + d
	Total	a + c	b + d	N = a+b+c+d

The conventional chi-square statistic for this 2 x 2 table is (Conover, 1971)

$$T = \frac{N(ad - bc)^2}{(a+b)(c+d)(a+c)(b+d)} \quad \ldots\ldots\ldots\ldots\ldots\ldots (7.1)$$

Considering the dependent variable as a zero-one variable, we have in the notation of part I of this thesis (sections 1.4 and 2.1) :-

The mean of the dependent variable:

$\bar{X} = (b + d) / N$

/87...

The variance of the dependent variable:

$$S^2 = ((b + d) - N\bar{X}^2) / N$$
$$= ((b + d) - N(b + d)^2/N^2) / N$$
$$= (b + d)(1 - (b + d)/N) / N$$
$$= (b + d)(N - (b + d)) / N^2$$
$$= (b + d)(a + c) / N^2$$

Number of observations in the first group:

$$M_1 = a + b$$

Number of observations in the second group:

$$M_2 = c + d$$

The mean of the first group:

$$\bar{Y}_1 = b/(a + b)$$

The mean of the second group:

$$\bar{Y}_2 = d/(c + d)$$

Hence from equation 2.1 the test statistic is

$$K = \frac{M_1 M_2}{N} \cdot \frac{(\bar{Y}_1 - \bar{Y}_2)^2}{S^2}$$
$$= \frac{(a + b)(c + d)}{N} \cdot \frac{(\frac{d}{c+d} - \frac{b}{a+b})^2}{(b+d)(a+c)/N^2}$$
$$= \frac{(a + b)(c + d)N^2}{N(b + d)(a + c)} \cdot \frac{(d(a + b) - b(c + d))^2}{((a + b)(c + d))^2}$$
$$= \frac{N(ad - bc)^2}{(b + d)(a + c)(a + b)(c + d)} \quad \ldots\ldots\ldots\ldots (7.2)$$
$$= T \text{ from (7.1)}$$

/88...

Thus the test statistics are identical in both types of A.I.D., and if the predictor is also dichotomous then T naturally has a chi-square distribution with one degree of freedom since it is derived from a 2 x 2 contingency table, and we have already shown in section 2.1 that K is likewise distributed in this case.

7.1.2 Some properties of the technique.

Since the merging process compares two rows at a time this implies that, if they have the same proportional distribution on each category of the dependent variable, these two rows will be merged of necessity. Consider two such rows depicted in figure 7.4 in which $a_k/A = b_k/B = C_k$ say, for all k.

Figure 7.4. A subtable of two rows.

		Dependent variable						
		1	2	3	..	..	d	Total
Predictor	Row i	a_1	a_2	a_3	..	..	a_d	A
	Row j	b_1	b_2	b_3	..	..	b_d	B
	Total	n_1	n_2	n_3	..	..	n_d	n

The chi-square for this subtable is calculated by

$$X = \sum_{k=1}^{d} \left(\frac{(a_k - n_k A/n)^2}{n_k A/n} + \frac{(b_k - n_k B/n)^2}{n_k B/n} \right) \quad \ldots\ldots\ldots (7.3)$$

since the expected value in each cell is given by the product of the marginal totals divided by the grand total.

/89...

Thus the test statistics are identical in both types of A.I.D., and if the predictor is also dichotomous then T naturally has a chi-square distribution with one degree of freedom since it is derived from a 2 x 2 contingency table, and we have already shown in section 2.1 that K is likewise distributed in this case.

7.1.2 Some properties of the technique.

Since the merging process compares two rows at a time this implies that, if they have the same proportional distribution on each category of the dependent variable, these two rows will be merged of necessity. Consider two such rows depicted in figure 7.4 in which $a_k/A = b_k/B = C_k$ say, for all k.

Figure 7.4. A subtable of two rows.

		Dependent variable						
		1	2	3	..	..	d	Total
Predictor	Row i	a_1	a_2	a_3	..	..	a_d	A
	Row j	b_1	b_2	b_3	..	..	b_d	B
	Total	n_1	n_2	n_3	..	..	n_d	n

The chi-square for this subtable is calculated by

$$X = \sum_{k=1}^{d} \left(\frac{(a_k - n_k A/n)^2}{n_k A/n} + \frac{(b_k - n_k B/n)^2}{n_k B/n} \right) \quad \ldots\ldots\ldots (7.3)$$

since the expected value in each cell is given by the product of the marginal totals divided by the grand total.

/89...

Since $a_k = AC_k$ and $b_k = BC_k$ it follows that

$$n_k = a_k + b_k$$
$$= AC_k + BC_k$$
$$= (A + B)C_k$$
$$= nC_k$$

So that

$$a_k - n_k A/n = AC_k - nC_k A/n = 0$$

and

$$b_k - n_k B/n = BC_k - nC_k B/n = 0$$

So that each numerator in the sum (7.3) is zero, and hence in this case $X = 0$, which is its minimum value and thus always not significant.

It is perhaps not well known, since this result does not appear in the literature, that standard A.I.D. duplicates this result by never splitting a predictor between two categories with identical means. This is demonstrated by considering a four category predictor with the middle two categories having the same mean, and showing that a split between these two categories can never be better than both the other splits available. Predictors with more categories are included in this result since the shift of the split point by one place is unaffected by the precise nature of the "outside" categories - whether they consist of a single category or more, their contribution to the test statistic remains unaltered. Although it appears as if only the monotonic case is being considered, this monotonic variable is arbitrary and hence the result for the free case (and subsequently the "floating"

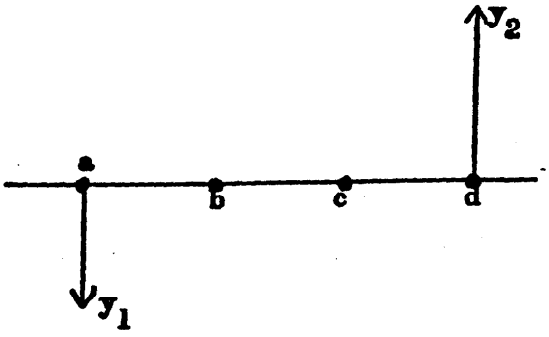
/90...

case, section 7.2) is also implied.

Since the test statistic is unaffected by a change in location of the dependent variable, take the middle two categories as having zero mean in order to simplify the algebra.

Let the sample sizes in each of the four groups be a, b, c and d respectively, each of these quantities being strictly positive. Let the mean of the first category be y_1, and of the last category y_2 where y_1 and y_2 are unconstrained. The four categories are depicted in figure 7.5.

Figure 7.5. Four category example whose middle two categories have identical means.



The between-sum-of-squares is given by

$$\text{B.S.S.} = n_1(\bar{X}_1 - \bar{X})^2 + n_2(\bar{X}_2 - \bar{X})^2 \quad \ldots\ldots\ldots (7.4)$$

where n_i and $\bar{X}_i$ are the number of observations and their mean in eacn part of the split and

$\bar{X}$ = the grand mean = $(n_1\bar{X}_1 + n_2\bar{X}_2) / (n_1 + n_2)$

$\ldots\ldots\ldots\ldots (7.5)$

/91...

So

$$\text{B.S.S.} = n_1\bar{X}_1^2 - 2n_1\bar{X}_1\bar{X} + n_1\bar{X}^2 + n_2\bar{X}_2^2 - 2n_2\bar{X}_2\bar{X} + n_2\bar{X}^2$$

$$= n_1\bar{X}_1^2 + n_2\bar{X}_2^2 - 2\bar{X}(n_1\bar{X}_1 + n_2\bar{X}_2) + (n_1 + n_2)\bar{X}^2$$

$$= n_1\bar{X}_1^2 + n_2\bar{X}_2^2 - 2\bar{X}(n_1 + n_2)\bar{X} + (n_1 + n_2)\bar{X}^2 \quad \text{from (7.5)}$$

$$= n_1\bar{X}_1^2 + n_2\bar{X}_2^2 - (n_1 + n_2)\bar{X}^2 \quad \ldots\ldots\ldots\ldots (7.6)$$

Note that the last term on the right of equation (7.6) is independent of how the group is split since $n_1 + n_2$ is the total sample size and $\bar{X}$ the grand mean, both of which remain unchanged.

The proof will proceed by showing that if the split in the middle yields a larger B.S.S. than both the other two possible splits, then a contradiction results.

If the split is after the first category, then in the above notation:

$n_1 = a, \quad n_2 = b+c+d, \quad \bar{X}_1 = y_1, \quad \bar{X}_2 = dy_2/(b+c+d),$

and hence from (7.6)

$$(\text{B.S.S.})_1 = ay_1^2 + d^2y_2^2/(b+c+d) - (n_1 + n_2)\bar{X}^2 \quad \ldots\ldots\ldots (7.7a)$$

If the split is after the second category:

$n_1 = a+b, \quad n_2 = c+d, \quad \bar{X}_1 = ay_1/(a+b), \quad \bar{X}_2 = dy_2/(c+d),$

and hence from (7.6)

$$(\text{B.S.S.})_2 = a^2y_1^2/(a+b) + d^2y_2^2/(c+d) - (n_1 + n_2)\bar{X}^2 \quad \ldots\ldots\ldots (7.7b)$$

Finally if the split is after the third category:

$n_1 = a+b+c, \quad n_2 = d, \quad \bar{X}_1 = ay_1/(a+b+c), \quad \bar{X}_2 = y_2,$

/92...

and hence from (7.6)

$$(B.S.S.)_3 = a^2y_1^2/(a+b+c) + dy_2^2 - (n_1 + n_2)\bar{X}^2 \quad \ldots\ldots\ldots\ldots (7.7c)$$

If the split after the second category dominates both other splits, then

$$(B.S.S.)_1 < (B.S.S.)_2 \quad \text{and} \quad (B.S.S.)_2 > (B.S.S.)_3$$

That is

$$ay_1^2 + d^2y_2^2/(b+c+d) < a^2y_1^2/(a+b) + d^2y_2^2/(c+d) \quad \ldots\ldots\ldots (7.8a)$$

and

$$a^2y_1^2/(a+b) + d^2y_2^2/(c+d) > a^2y_1^2/(a+b+c) + dy_2^2 \quad \ldots\ldots\ldots (7.8b)$$

These equations give respectively

$$ay_1^2/(a+b) < d^2y_2^2/((c+d)(b+c+d)) \quad \text{for } b \neq 0 \quad \ldots\ldots\ldots (7.9a)$$

and

$$a^2y_1^2/((a+b)(a+b+c)) > dy_2^2/(c+d) \quad \text{for } c \neq 0 \quad \ldots\ldots\ldots (7.9b)$$

and hence

$$ay_1^2(c+d) / (dy_2^2(a+b)) < d/(b+c+d) \quad \ldots\ldots\ldots (7.10a)$$

and

$$ay_1^2(c+d) / (dy_2^2(a+b)) > (a+b+c)/a \quad \ldots\ldots\ldots (7.10b)$$

for $d > 0$, $y_2 \neq 0$, $b > 0$, $a > 0$.

The equations (7.10) together imply

$$(a+b+c)/a < d/(b+c+d)$$

which in turn implies

$$(a+b+c)(b+c+d) < ad$$

That is

$$a(b+c) + ad + (b+c)^2 + d(b+c) < ad$$

ie,

$$(b+c)(a+b+c+d) < 0 \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (7.11)$$

Clearly (7.11) is impossible since the sample size within each category is non-negative. Hence there must exist a better split than that separating two identical contiguous categories. A similar proof to the above shows that this result also holds in the three category case.

7.1.3 Convergence of the procedure.

This subsection demonstrates that the algorithm that has been presented above (section 7.1) for analysing a contingency table will converge to a stable solution in a finite number of steps when there are an infinite number of observations per cell provided that the significance level for splitting is smaller than that for merging. This theoretical argument will then provide some sufficient conditions for convergence of the algorithm in the case of finite samples.

Asymptotically it can be shown (see Kendall and Stuart, 1961, Vol 2, page 577) that the total chi-square for a contingency table can be partitioned into a component for a specified subtable and a component for the resulting table after the subtable has been "collapsed". In particular, referring to figure 7.1 presented earlier, let the total chi-square for

/94...

this table be denoted by T (with (c-1)(d-1) degrees of freedom). Let the statistic for merging any two rows i and j (merging of course over all d columns) be denoted by X (with d-1 degrees of freedom). Given that these two rows i and j are merged let the resultant (c-1) x d table have chi-square Y (with (c-2)(d-1) degrees of freedom). Hence the asymptotic result quoted above may be written as

$$T = X + Y \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (7.12)$$

Each time two rows are merged a new table results, and the process is repeated with the T of the new (reduced) table taking the value of the Y of the previous table.

For any merge, X is non-negative and hence by equation (7.12) the total chi-square for the succession of tables obtained by merging various pairs of rows is a non-increasing sequence $T_0, T_1, T_2, \ldots$, say. Any subtable obtained from the original c x d table, must have a non-negative chi-square T_i that cannot exceed the original chi-square T_0, and hence is bounded. While the number of possible subtables may be large, there is a finite number of them, and hence in any particular case T_i can only assume a finite number of values.

The procedure of merging and splitting produces a series of bounded values $\{T_i\}$ where $T_i \leqslant T_{i-1}$ if a merge occurs and $T_i \geqslant T_{i-1}$ if a split occurs. Because there is a finite number of subtables, the procedure will fail to

/95...

this table be denoted by T (with (c-1)(d-1) degrees of freedom). Let the statistic for merging any two rows i and j (merging of course over all d columns) be denoted by X (with d-1 degrees of freedom). Given that these two rows i and j are merged let the resultant (c-1) x d table have chi-square Y (with (c-2)(d-1) degrees of freedom). Hence the asymptotic result quoted above may be written as

$$T = X + Y \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (7.12)$$

Each time two rows are merged a new table results, and the process is repeated with the T of the new (reduced) table taking the value of the Y of the previous table.

For any merge, X is non-negative and hence by equation (7.12) the total chi-square for the succession of tables obtained by merging various pairs of rows is a non-increasing sequence $T_0, T_1, T_2, \ldots$, say. Any subtable obtained from the original c x d table, must have a non-negative chi-square T_i that cannot exceed the original chi-square T_0, and hence is bounded. While the number of possible subtables may be large, there is a finite number of them, and hence in any particular case T_i can only assume a finite number of values.

The procedure of merging and splitting produces a series of bounded values $\{T_i\}$ where $T_i \leqslant T_{i-1}$ if a merge occurs and $T_i > T_{i-1}$ if a split occurs. Because there is a finite number of subtables, the procedure will fail to

/95...

converge only if it "gets into a loop", that is a particular configuration (subtable) occurs and recurs after a series of merges and splits. If the criterion for splitting is more stringent, that is demands a higher critical value, than that for merging as suggested in the algorithm, then such "looping" is impossible.

To see this, first note that a loop, if it exists, must contain the same number of merges as splits in order to return to an identical point in the analysis. Consider the series $\{T_i\}$ for such a loop. By assumption the T value will be identical at the beginning and end of the loop since the same subtable has been attained. However, because of the merging criterion, each merge can reduce a particular T_i by at most some critical value k, say. On the other hand a split will increase a particular T_i by at least some critical value k', say, where the algorithm demands $k' > k$. If there are m merges and hence also m splits in the "loop" the starting T value of the "loop" will be decreased by at most mk for the merges, and increased by at least $mk' > mk$ for the splits. Hence a contradiction results since the value for T at the end of the "loop" cannot coincide with the value at the beginning of the "loop". So it must be concluded that such "looping" is impossible.

For finite samples, the relationship $T = X + Y$ is only true approximately (except in some special cases to be considered in a moment) but Kendall and Stuart (1961, Vol 2, page 577)

/96...

claim "the approximate partition is good enough for most practical purposes."

We briefly examine some details of the approximation (7.12) $T = X + Y$. Let the two rows that are merged be given by figure 7.4 (in subsection 7.1.2 above) so that

$$X = \sum_{k=1}^{d} \left(\frac{(a_k - An_k/n)^2}{An_k/n} + \frac{(b_k - Bn_k/n)^2}{Bn_k/n} \right) \quad \text{from (7.3)}$$

$$= \frac{1}{nAB} \sum_{k=1}^{d} \frac{1}{n_k} \left(n^2(a_k^2 B + b_k^2 A) - 2n_k nAB(a_k + b_k) + n_k^2 AB(A+B) \right)$$

$$= \frac{1}{nAB} \sum_{k=1}^{d} \frac{1}{n_k} \left(n^2(a_k^2 B + b_k^2 A) - n_k^2 nAB \right)$$

$$= \frac{1}{AB} \sum_{k=1}^{d} \frac{1}{n_k} \left((A+B)(a_k^2 B + b_k^2 A) - (a_k + b_k)^2 AB \right)$$

$$= \frac{1}{AB} \sum_{k=1}^{d} \frac{1}{n_k} (a_k^2 AB + b_k^2 A^2 + a_k^2 B^2 + b_k^2 AB - a_k^2 AB - b_k^2 AB - 2a_k b_k AB)$$

$$= \frac{1}{AB} \sum_{k=1}^{d} \frac{1}{n_k} (a_k B - b_k A)^2 \quad \ldots\ldots\ldots\ldots\ldots (7.13)$$

If the full table with chi-square T has column totals given by N_i $(i = 1, 2, \ldots, d)$ and $N = \sum N_i$, then the sub-table obtained from the total table by merging row i and row j is given by

Y = T − (contribution to chi-square from individual rows i & j) + (contribution to chi-square from the merged rows i & j)

That is

$$Y = T - \sum_{k=1}^{d} \left(\frac{(a_k - AN_k/N)^2}{AN_k/N} + \frac{(b_k - BN_k/N)^2}{BN_k/N} \right) + \sum_{k=1}^{d} \frac{(n_k - nN_k/N)^2}{nN_k/N}$$

$$= T - Z \quad \text{say.}$$

/97...

So,

$$Z = \frac{1}{ABnN}\sum_{k=1}^{d}\frac{1}{N_k}\left[Bn(Na_k-AN_k)^2 + An(Nb_k-BN_k)^2 - AB(Nn_k-nN_k)^2\right]$$

$$= \frac{1}{ABnN}\sum_{k=1}^{d} N_k^{-1}\,(N^2a_k^2Bn + A^2N_k^2Bn - 2a_kNN_kABn + N^2b_k^2An + B^2N_k^2An - 2b_kNN_kABn - N^2(a_k+b_k)^2AB - (A+B)^2N_k^2AB + 2(a_k+b_k)NN_knAB)$$

$$= \frac{1}{ABnN}\sum_{k=1}^{d} N_k^{-1}\,(N^2a_k^2AB + N^2a_k^2B^2 + N_k^2A^3B + N_k^2A^2B^2 + N^2b_k^2A^2 + N^2b_k^2AB + N_k^2A^2B^2 + N_k^2AB^3 - N^2a_k^2AB - N^2b_k^2AB - 2N^2a_kb_kAB - N_k^2A^3B - N_k^2AB^3 - 2N_k^2A^2B^2)$$

$$= \frac{1}{ABnN}\sum_{k=1}^{d} N_k^{-1}\,(N^2a_k^2B^2 + N^2b_k^2A^2 - 2N^2a_kb_kAB)$$

$$= \frac{N}{ABn}\sum_{k=1}^{d} N_k^{-1}\,(a_kB - b_kA)^2 \qquad \ldots\ldots\ldots\ldots\ldots\ldots (7.14)$$

The error in the approximation $T = X + Y = X + (T-Z)$ is obtained from equations (7.13) and (7.14) as

$$Z - X = \frac{1}{AB}\sum_{k=1}^{d}(a_kB - b_kA)^2\left(\frac{N}{nN_k} - \frac{1}{n_k}\right)$$

$$= \frac{1}{AB}\sum_{k=1}^{d}(a_kB - b_kA)^2\left(\frac{n_kN - nN_k}{nN_kn_k}\right) \qquad \ldots\ldots\ldots (7.15)$$

Two special cases when this error is zero are immediately apparent. Firstly, if the two rows i and j each have

/98...

the same proportion of observations within each category, chi-square X will be zero, which from (7.13) implies that $a_k B = b_k A$ for all k, and hence Z - X will be zero from (7.15). The second possibility is that $n_k N = n N_k$ for all k in (7.15) which implies that $n_k/N_k = n/N$, that is the column totals for the 2 x d subtable and the column totals for the original c x d table are in constant proportion.

Unfortunately, the error Z - X is not of constant sign for all contingency tables as evidenced by the examples in figure 7.6.

Figure 7.6. Two examples that demonstrate that the error of the approximation T = X+Y (see text) is not of constant sign.

A.

(a) Complete table

rows	1	3	9
	2	3	21
	3	30	6

T = 32.5

(b) Subtable

rows	1	3	9
	2	3	21

X = .9

(c) Subtable with rows 1 & 2 merged.

rows	1&2	6	30
	3	30	6

Y = 32.0

Z = T-Y = .5 < .9 = X

ie. Z-X = -0.4 < 0

B.

(a) Complete table

rows	1	15	5
	2	5	15
	3	30	20

T = 11.025

(b) Subtable

rows	1	15	5
	2	5	15

X = 10.000

(c) Subtable with rows 1 & 2 merged.

rows	1&2	20	20
	3	30	20

Y = .9

Z = T-Y = 10.125 > 10.000 = X

ie. Z-X = .125 > 0

/99...

Note that in the two examples of figure 7.6 the error is comparatively small verifying the observation of Kendall and Stuart (1961) mentioned above.

The argument to show that the algorithm converges in the asymptotic case, extends to the finite case as long as the difference between the merging and splitting criteria is greater than the maximum possible error that could arise using the approximation $T = X+Y$ for the particular data under consideration. In practice we have performed many analyses and simulations using the critical value for the 5% significance level as the merging criterion and the critical value for the 4.9% significance level as the splitting criterion without any problems. The computer program provides a safeguard against a possible infelicitous choice of criteria that results in "looping" by limiting the total number of mergings and splittings within the analysis on any one predictor.

7.1.4 An alternative method (with an inherent weakness).

Since it is the aim of the procedure to "reduce" the original contingency table to the most significant subtable, it may be thought that the intermediate criterion on which the merging and splitting process is based, should be the significance of the whole table rather than that of just the subset of the table that is being considered for merging or splitting.

/100...

These two procedures are identical only to the degree of the approximation discussed in the previous sub-section. A convenient reference to the exact theory is that of Maxwell (1961, page 52ff) who cites the original references and gives examples.

An apparent advantage in using the significance of the total table at each stage of the analysis, is that the Bonferroni estimate of the true significance could be incorporated at each stage. Unfortunately this leads to an undesirable property illustrated by the example in figure 7.7.

Figure 7.7. Example leading to a contradiction.

		Dependent variable	
		1	2
Predictor	Row 1	a	b
	Row 2	a	b
	Row 3	c	d
	Row 4	e	f

This 4 x 2 table has row 1 identical to row 2, the actual numbers in each cell being purely arbitrary. The test statistic will be chi-square distributed with three degrees of freedom and we may choose the test statistic to have a value of T = 7.82 so that it is just significant at the 5% level. Combining the first two rows which is expected of the method since they are identical (see section 7.1.2), reduces the table to a 3 x 2 contingency table with unchanged test sta-

/101...

tistic value $T = 7.82$ in this case. This value is significant at the 2% level. However, the Bonferroni multiplier for a monotonic predictor is 3 (the number of 3×2 tables derivable from the original 4×2 table), which gives significance at the $3 \times 2\% = 6\%$ level, that is no longer is there significance at the 5% level. The situation for a free predictor is even worse since the number of derivable subtables is $\binom{4}{2} = 6$ which raises the significance level to $6 \times 2\% = 12\%$.

Thus, while the use of the Bonferroni estimate of the significance level at each stage of the merging process is intuitively appealing, this method would lead to an undesirable conclusion as rows 1 and 2 would not be merged.

7.2 Types of predictor.

This section considers three types of predictors: monotonic and free which are used in conventional A.I.D., and a predictor that we have dubbed "floating" which could well be used in A.I.D., but is discussed in this section only in terms of CHAID. A generalization of the "floating" predictor which includes the types monotonic and free as special cases, will be discussed briefly.

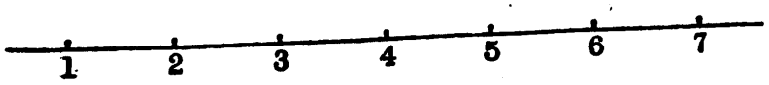
Each type of predictor will be discussed in the context of CHAID, examples given to show that the "splitting" process is necessary, the formula for the Bonferroni multipliers derived, and some special considerations in the analysis that affect a particular type of predictor will be noted where appropriate.

/102...

7.2.1 Monotonic predictors.

As in A.I.D., a monotonic predictor is one whose categories lie on an ordinal scale. This may be depicted in figure 7.8 which indicates that the categories may be conceived as laid out on a line.

Figure 7.8. Example of a monotonic predictor with seven categories.



At each stage in the merging process, only adjacent categories may be merged, or a compound group merged with an adjacent category or group. The splitting process is identical to A.I.D., except that only compound groups are considered for splitting, and not the total set of categories (unless, of course, all the categories have been merged down to a single compound group).

To show that splitting a compound group may arise in the case of a monotonic predictor consider the example in figure 7.9. (See next page).

The analysis would proceed from the full table depicted in (a). Of the three possible merges, 1&2, 2&3, and 3&4, that of 1&2 is the least significant, and is depicted in (b). After this merge has taken place, there are two possible merges:

/103...

Figure 7.9. Example of splitting a monotonic predictor.

(a) Full table

		Dependent variable 1	2
Predictor	Row 1	120	80
	Row 2	70	65
	Row 3	25	40
	Row 4	140	130

(b) Least significant two rows

		Dependent variable 1	2	
Predictor	Row 1	120	80	$\chi^2 = 2.180$
	Row 2	70	65	

(c) Next least significant difference

		Dependent variable 1	2	
Predictor	Row 3	25	40	$\chi^2 = 3.758$
	Row 4	140	130	

(d) Reduced table

		Dependent variable 1	2	
Predictor	Rows 1&2	190	145	$\chi^2 = 3.745$
	Rows 3&4	165	170	

(e) Best split of completely merged table

		Dependent variable 1	2	
Predictor	Row 1	120	80	$\chi^2 = 5.632$
	Rows 2,3&4	235	235	

(1&2)&3 and 3&4. The least significant is 3&4 which is depicted in (c). The table has now been reduced to (d) which has a chi-square value of 3.745, so that the full table would be collapsed into a single row if the critical value of 3.84 (corresponding to the 5% significance level) were used.

/104...

If the splitting stage were unnecessary, the maximum chi-square value attainable by a split would be the same as that obtained by the last merge, ie. 3.745. In this case however, the split depicted in (e) is better, yielding a chi-square value of 5.632 which is significant even at the 2% level.

The formula for the appropriate Bonferroni multiplier follows as a simple extension to that obtained for conventional A.I.D. as in section 4.3. In general, the number of ways of partitioning a c-category monotonic predictor into r groups is identical to the number of ways of placing r-1 divisions in the c-1 gaps between the categories in figure 7.8. Hence the Bonferroni multiplier is given by the binomial coefficient

$$B_{monotonic} = \binom{c-1}{r-1} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (7.16)$$

Note, that in particular when r=2, a simple binary split, the Bonferroni multiplier is (c-1).

This coefficient is easily calculated on a computer, and is produced routinely in program CHAID. Tables of the binomial coefficients are readily available, see for example Hald (1952).

7.2.2 Free predictors.

Again, as in conventional A.I.D., a free predictor is one whose categories can be considered purely nominal. At each stage in the merging process any two categories or compound categories may be combined. Similarly, during the splitting phase, a compound group may be divided with no restriction as to the grouping of the categories.

/105...

To show that splitting a compound group may arise in the case of a free predictor, consider the example in figure 7.10.

Figure 7.10. Example of splitting a free predictor.

(a) Full table

		Dependent variable 1	2
Predictor	Row 1	69	71
	Row 2	113	96
	Row 3	59	82
	Row 4	32	37

(b) Least significant two rows

		Dependent variable 1	2	
Predictor	Row 1	69	71	χ^2 = .157
	Row 4	32	37	

(c) Next least significant difference

		Dependent variable 1	2	
Predictor	Rows 1&4	101	108	χ^2 = 1.379
	Row 2	113	96	

(d) Reduced table

		Dependent variable 1	2	
Predictor	Rows 1,2&4	214	204	χ^2 = 3.691
	Row 3	59	82	

(e) Best split of completely merged table

		Dependent variable 1	2	
Predictor	Rows 1&2	182	167	χ^2 = 4.078
	Rows 3&4	91	119	

/106...

The analysis starts with the full table (a) and finds that the most insignificantly different pair of rows is 1 and 4, with a chi-square of .157 as in (b). Combining rows 1 and 4, this compound row and rows 2 and 3 are analysed the least significantly different being the compound 1&4 and the original row 2 as in (c). The compound 1,2&4 has a chi-square of 3.691 when compared with row 3 as shown in (d), so that the full table can be collapsed into a single row if the critical value 3.84 corresponding to the 5% significance level is chosen.

If the splitting stage were unnecessary for a free predictor, the maximum chi-square attainable by splitting the compound row with all categories, would again be 3.691. However, as seen in (e), splitting the first two rows from the last two gives rise to a chi-square of 4.078 which is significant at the 5% level and would cause the program to split the group at this point. (Note, however, that in both this case and the monotonic example of figure 7.9, the 2 x 2 subtables given in (e) would not be considered significant after the final Bonferroni test.

The formula for the appropriate Bonferroni multiplier is not trivial to obtain, but fortunately has been derived as a special case by Feller (1968) when considering the classical occupancy problem. In the language of this problem the c categories of the predictor can be conceived as being c balls to be placed in r cells where each cell represents

/107...

a group into which a category is placed. Feller (1968) shows on page 102 that the probability of such an assignment such that all cells are occupied is

$$p_0(c,r) = \sum_{i=0}^{r-1} (-1)^i \binom{r}{i} (1 - \tfrac{i}{r})^c$$

where his notation has been modified to concur with ours, and the last term of the summation (when i=r) which is zero, has been removed. Since this probability has been calculated over all the r^c possible arrangements of balls (categories) in cells (groups), and it is immaterial in our case which one of the $r!$ permutations of cells (groups) are used, the number of ways of partitioning the categories into r groups is

$$B_{free} = r^c p_0(c,r) \,/\, r!$$

$$= \sum_{i=0}^{r-1} (-1)^i \binom{r}{i} (1 - \tfrac{i}{r})^c \; \frac{r^c}{r!}$$

$$= \sum_{i=0}^{r-1} (-1)^i \; \frac{r!}{i!(r-i)!} \cdot \frac{(r-i)^c}{r^c} \cdot \frac{r^c}{r!}$$

$$= \sum_{i=0}^{r-1} (-1)^i \; \frac{(r-i)^c}{i!(r-i)!} \qquad \ldots\ldots\ldots\ldots\ldots\ldots (7.17)$$

When only two groups are formed, ie. r=2, then (7.17) reduces to

$$\sum_{i=0}^{1} (-1)^i \; \frac{(2-i)^c}{i!(2-i)!} = 2^{c-1} - 1$$

the result obtained in part I of this thesis in section 1.1.

/108...

Equation (7.17) is easily evaluated on a computer, and is produced routinely in program CHAID. Some commonly used values are given in table 7.1 below for the edification of the reader.

Table 7.1. Bonferroni multipliers in the free case.

Categories c	Number of groups, r 2	3	4	5	6
2	1				
3	3	1			
4	7	6	1		
5	15	25	10	1	
6	31	90	65	15	1
7	63	301	350	140	21
8	127	966	1701	1050	266
9	255	3025	7770	6951	2646
10	511	9330	34105	42525	22827
11	1023	28501	145750	246730	179487
12	2047	86526	611501	1379400	1323652
13	4095	261625	2532530	7508501	9321312
14	8191	788970	10391745	40075035	63436373
15	16383	2375101	42355950	210766920	420693273
16	32767	7141686	171798901	1096190550	2734926558

A problem peculiar to a free predictor is the labour involved during the splitting phase to determine the optimum binary split. In conventional A.I.D. this problem is simplified by the method adopted by Sonquist and Morgan (1964) who show that the optimum split can be found by ordering the categories according to their mean on the dependent variable and then treating these ordered categories in the monotonic fashion. This technique carries over to the case of a dichotomous dependent variable where the categories can be ordered according to

the proportions that each category contains of one of the dichotomies. In general, however, for a non-dichotomous (and hence non-ordered) dependent variable, the categories cannot be so ordered. Consider as an example the data in figure 7.11 where the number of observations in each cell could also be considered as row percentages.

Figure 7.11. Example of a free predictor whose categories cannot be ordered.

		Dependent variable			
		1	2	3	4
Predictor	Row 1	10	40	20	30
	Row 2	20	30	40	10
	Row 3	30	20	10	40
	Row 4	40	10	30	20

The rows are in fact ordered according to both category 1 and category 2 of the dependent variable. However, they are clearly not in order according to the last two categories of the dependent variable. Further, any permuting of the rows, while perhaps improving the ordering in the last two columns, will cause a deterioration of the perfect ordering in the first two.

At this stage we have been unable to find any general method of ordering the rows such that treating them in a monotonic fashion will still guarantee finding the best binary split. Since no better method can be found, CHAID analyses all the possible splits to find the best one. As this can be time

/110...

consuming, even for a computer, when a compound group contains a large number of categories, CHAID offers an option that inhibits the splitting phase for any particular predictor. In practice ten-category predictors have been analysed including the splitting phase, but it is felt that with more than ten categories the impact of these extra calculations may be felt. For further details on timing, see chapter 8.

7.2.3 Floating predictors.

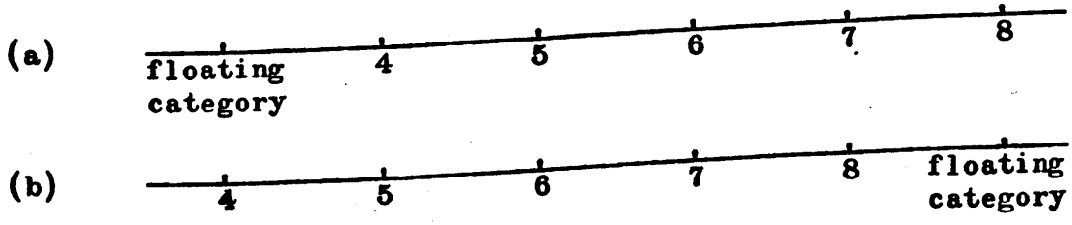
In many cases of practical interest, the categories of a predictor lie on an ordinal scale with the exception of a single category that either does not belong with the rest, or whose position on the ordinal scale is unknown. This situation typically arises when an investigation allows for an "Unknown" or "Do not know" category.

Missing information is often recoded by some technique such as by Preece (1971) who allocates a possible score or value to the missing information; however, since such missing data does not upset A.I.D. type analyses it is usual to assign a special category to them, see for example Morgan and Messenger (1973). Thus, whilst the other categories may be considered monotonic, this category of missing information is unplaced in the ordered set and we may call it a "floating" category, an attribute that we pass on to the predictor - a "floating" predictor.

/111...

The program's treatment of a floating predictor is as follows. During the merging phase the other categories may be combined as described in section 7.2.1 - the monotonic predictor, and the floating category may be merged with any one of the other categories or compound row. Once the floating category has been combined with another category or group, the merging process treats the predictor as if it were completely monotonic - the floating category having now been placed in the ordered scale. The splitting phase treats any compound group that does not contain the floating category as a monotonic set. A compound group containing the floating category is analysed for splitting as if it were monotonic but the floating category may appear on either side of the split point. This is accomplished by taking the monotonic categories and first placing the floating category on the one end of the ordinal scale, and then on the other. Diagrammatically, this can be depicted as in the example of figure 7.12 where the compound group contains five monotonic categories and the floating category.

Figure 7.12. Example of a compound group containing the floating category.



/112...

The compound group is first analysed for splitting as if it were a monotonic group with the ordering given in (a), and then with the ordering given in (b). The last possible split point in (b), that between category 8 and the floating category, is not actually considered since it gives rise to the same split as locating the split point between the floating category and category 4 in (a).

The formula for the Bonferroni multiplier comes from a simple extension of the monotonic case. A c-category floating predictor partitioned into r groups will either have the floating category in a group by itself, or combined with some other category or categories. In the former case, the r groups are formed by the single floating category and partitioning the remaining c-1 monotonic categories into r-1 groups. From (7.16) this can be done in

$\binom{c-2}{r-2}$ ways.

When the single floating category does not appear by itself, the c-1 monotonic categories may be divided into r groups and the floating category assigned to one of these r groups. This results in

$r\binom{c-2}{r-1}$ possible partitions.

Hence, the Bonferroni multiplier for the floating predictor is given by

/113...

$$B_{float} = \binom{c-2}{r-2} + r\binom{c-2}{r-1}$$

$$= \frac{(c-2)!}{(r-2)!(c-r)!} + \frac{r(c-2)!}{(r-1)!(c-r-1)!}$$

$$= \frac{(c-2)!}{(r-1)!(c-r)!}\ (\ r-1 + r(c-r)\ )$$

$$= \binom{c-1}{r-1}\frac{r-1+r(c-r)}{c-1}$$

$$= \frac{r-1+r(c-r)}{c-1}\ B_{monotonic} \quad \ldots\ldots\ldots\ldots (7.18)$$

This formula is easily evaluated on a computer, and values are produced routinely in program CHAID. Some commonly used values are given in table 7.2.

Table 7.2. Bonferroni multipliers in the floating case.

Categories c	Number of groups, r				
	2	3	4	5	6
2	1				
3	3	1			
4	5	5	1		
5	7	12	7	1	
6	9	22	22	9	1
7	11	35	50	35	11
8	13	51	95	95	51
9	15	70	161	210	161
10	17	92	252	406	406
11	19	117	372	714	882
12	21	145	525	1170	1722
13	23	176	715	1815	3102
14	25	210	946	2695	5247
15	27	247	1222	3861	8437
16	29	287	1547	5369	13013

/114...

7.2.4 Interleaved-monotonic predictors.

The above section considers only a single floating category that cannot be placed on the overall monotonic scale. It is easy to conceive of situations where two or more such floating categories exist. For example, a respondent to a questionnaire refuses to answer a particular question and either gives no reason for his refusal, or gives a reason which has pertinent information.

Consider a generalization of this problem where the floating categories themselves may be placed on a monotonic scale. A typical situation that may arise in educational research is where a respondent is required to state his highest educational qualification. Within any one country these answers may be placed on a monotonic scale. However, a "school leaving certificate" in one country may not be at the same academic level as an apparently equivalent one from another country. If some objective investigation could determine an accurate correspondence between the two scales no problem would arise. Usually, no such correspondence exists, and the result is two (or more) monotonic scales interleaved in some unknown fashion. In this case, and others of a similar genre, we define an interleaved-monotonic scale as follows.

/115...

Start with a c-category predictor containing all the possible categories that may occur. Partition these categories into s non-overlapping subsets such that each subset forms an ordinal (monotonic) scale, the i^{th} subset containing c_i categories (i=1,2,..,s). Any partition that simultaneously retains the ordering within each subset is a valid partition for this interleaved-monotonic predictor. It is assumed that each subset is ordered in the same direction, though another type of predictor without this condition could be considered, but it is difficult to find a practical example where this would be useful.

If in the above example of educational qualifications, the people in the survey were drawn from one of two possible countries, then s=2 would result and c_1 would denote the number of levels of education available in the one country and c_2 the number in the other country. Note that a "missing information" category could also be provided by taking s=3 and c_3=1, this being the "missing information" category.

All the types of predictor discussed in previous sections are special cases of this interleaved-monotonic predictor. Taking s=1 and hence c_1=c gives the standard monotonic predictor. A free predictor arises when s=c and each c_i=1. The floating predictor has s=2 and either c_1 or c_2 equal to one (the floating category) the ordering of the c_i being immaterial.

/116...

The implementation of such a generalized predictor is not trivial, but special cases are more easily handled. Similarly, the calculation of the Bonferroni multipliers does not appear an easy task in general, but again for a particular case the problem is simplified. The idea of an interleaved-monotonic predictor is put forward as a possibly useful generalized predictor but it is temporarily shelved pending future investigation and research.

CHAPTER 8. Empirical Investigation.

8.1 Simulation under the null hypothesis.

In the previous two chapters a number of claims have been made regarding the efficacy of CHAID in practice. Specifically, one of the main aims was to control the type I error - the possibility of finding some "relationship" in the data, indicated by a split, which is fortuitous.

A simulation of size two hundred was made for each type of predictor. Each simulation generated 480 data elements which were partitioned into 2, 3, 5, 8 and 10 category predictors. The simulation extended over 2, 3, 4 and 5 category dependent variables, making ten trials for each combination of number of categories in the predictor and dependent variable.

The random data were obtained by generating 480 uniform random numbers u_i say, by a Tausworthe generator (appendix A discusses this and gives the relevant coding as part of the normal random number generator), and assigning the observation to category $[1 + du_i]$ for a d-category dependent variable. A c-category predictor was assigned cyclically 1,2,..,c,1,2,.. over the 480 data elements ensuring an equal allocation to each category.

In each case the merging criterion was significance at the 5% level, and the splitting criterion was set to 4.9%. The final result was only considered significant if the table

/118...

was not collapsed into a single row, and the resulting contingency table was significant at the 5% level after correction by its Bonferroni multiplier.

The results of the simulation are displayed in table 8.1 below. Under the null hypothesis, each cell is expected to have 0.5 trials significant, making the total for each row 2.5 and the total for each column 2.

Table 8.1. Number of significant results at 5% level in data simulating the null hypothesis.

(a) Monotonic predictors.

Dependent variable	Number of categories in predictor 2	3	5	8	10	Total
2 categories	0	1	1	0	0	2
3 categories	1	0	0	1	0	2
4 categories	0	1	1	2	0	4
5 categories	1	0	1	0	1	3
Total	2	2	3	3	1	11 / 200 = 5½%

(b) Free predictors.

Dependent variable	Number of categories in predictor 2	3	5	8	10	Total
2 categories	0	1	1	0	0	2
3 categories	1	0	0	1	0	2
4 categories	0	2	1	0	0	3
5 categories	1	0	0	0	0	1
Total	2	3	2	1	0	8 / 200 = 4%

(c) Floating predictors.

Dependent variable	Number of categories in predictor 2	3	5	8	10	Total
2 categories	1	0	0	0	0	1
3 categories	1	0	1	1	1	4
4 categories	1	0	1	1	1	4
5 categories	0	0	1	0	0	1
Total	3	0	3	2	2	10 / 200 = 5%

/119...

Each of the three sets of simulations had an approximately overall 5% error rate matching the 5% significance level that was used. The free predictor (b) had a 4% error rate which is not significantly different from the expected 5%, although one may feel intuitively that the large Bonferroni multipliers for a free predictor (see table 7.1) could cause the test to be too stringent.

Table 8.1 is a satisfactory result and should instil confidence in CHAID users that they are not being misled by chance relationships.

8.2 Timing.

Since the simulation covered a variety of dependent variables and predictors, the computer execution time for this experiment can provide an "average" execution time of the routines for each type of predictor. These timing estimates are based on the IBM S/370 Model 145 computer performing under IBM's VS1 operating system.

The average time for each routine is obtained from the total time of each simulation minus the overhead required to generate the data. This last quantity was obtained by re-running a simulation excluding the actual CHAID analysis. Hence the time for each routine includes the linkage time both to and

/120...

from the subroutines. Since the splitting process within a free predictor is time consuming, an additional simulation was run inhibiting the splitting process. The result of these runs and the calculations are given in table 8.2.

Table 8.2. Time (in seconds) for each routine on an IBM S/370 Model 145.

	Type of predictor.			
	Monotonic	Free	Free (no split)	Floating
	secs	secs	secs	secs
(a) Total time for 200 trials	148	559	206	158
(b) Overhead	23	23	23	23
(c) Net time (a) - (b)	125	536	183	135
(d) Time / predictor = (c)/200	0.63	2.7	0.91	0.68
(e) Number / second = 200/(c)	1.6	0.4	1.1	1.5

For those not conversant with the IBM S/370 Model 145, the "Gibson mix" index for this machine is 0.30 milliseconds. A discussion of computer "benchmarks" and a comparison with other machines is given by A.H.Duncan in the book edited by Yearsley and Graham (1973, chapter 5).

Table 8.2 indicates that a monotonic predictor is evaluated fastest, with the floating predictor only slightly slower. A free predictor takes about four times as long to analyse though this factor would be reduced if fewer categories per predictor were used. The extra time needed to analyse a free predictor is mainly due to the splitting process which has been

/121...

noted earlier to be cumbersome. Excluding the splitting process from the analysis of the free predictor reduces the time taken by a factor of about one third, bringing it more in line with the timings for the other types of predictor.

/122...

CHAPTER 9. An example.

The program CHAID has been used successfully in a number of analyses by the author. Some data with a dichotomous dependent variable, previously analysed by A.I.D., were re-examined with CHAID, yielding identical final groups although there were differences in some of the sub-analyses of predictors that were not used for subdividing the data.

For the purpose of describing the printout of CHAID and its interpretation, an analysis of student data at the University of the Witwatersrand is presented. The selection of students is based primarily on their matriculation examination record. There are a number of examination boards in South Africa, and there is a widespread belief that the standards set by the various boards are not equivalent. It has been said by various educationists that students who matriculate under some boards "do not perform as well as expected." This is partially borne out by a cross-tabulation of matriculation board and university performance, where students examined by a particular board consistently underperform in various faculties. Rather than immediately blame the examination system it was decided to investigate possible associations between the matriculation board and other background information available about students.

/123...

This investigation covered 669 students for the Certificate in the Theory of Accountancy (CTA) over the four years 1960 to 1963 inclusive. The matriculation boards were classified into four categories as follows.

(i) Joint Matriculation Board (JMB), covering 20% of the students.

(ii) Transvaal education boards including the Secondary School Certificate (TSSC) and the University Entrance Certificate (TUEC), covering 59% of the students.

(iii) National Senior Certificate (NSC), covering 14% of the students. (This was the main examination thought to be inferior).

(iv) Other boards, covering 7% of the students, consisting mainly of boards of other provinces and credentials presented by "overseas" students.

Seven "predictors" of the matriculation board were considered:

(i) Whether or not a student obtained a CTA within ten years of admission

(ii) Year of entry: 1960, 1961, 1962 or 1963

(iii) Number of commercial subjects included in matriculation curriculum (eg. Book-keeping, Mercantile Law, etc)

(iv) Number of non-academic subjects included in matriculation curriculum (eg. Typing, Woodwork, etc.)

(v) Type of school attended: "conventional" schools or "tutorial" colleges (called colloquially "cram" colleges)

(vi) Matriculation mathematics symbol ("A" to "F", "A" being a distinction, "F" a fail)

(vii) Matriculation English symbol ("A" to "F", "A" being a distinction, "F" a fail).

/124...

Sections of the analysis are displayed in appendix C, and are referenced by the circled numbers. Page (1) summarizes the information about the variables and the options used in the analysis.

Note that each category name is symbolized by a single printed character (in order to shorten later output). The first predictor of whether a student eventually obtained the CTA was inhibited from splitting the data (if it were significant enough - as it turned out it was not) as indicated by the word "NOT" in the column headed "USE?". The last two predictors were defined with their first category "?" as floating since this represented missing or unobtainable information.

The output of the analysis is split into three parts. The first from (2) to (7) summarizes the analysis of each group and subgroup. The second, from (8) to (11) gives the detail of the first part of the analysis within each predictor, showing how the categories were merged and split. Note that the similarity of the "missing information" category to each of the other categories is given and could be used to recode such information to the category most similar to it, for some non-CHAID processing. The final part, from (12) to (15) gives the cross tabulation between the dependent variable and each predictor for the total group, both before and after the merging process has taken place. Note that in practice, the second and third parts are given for each subgroup analysed by the program - this output has been omitted from appendix C to reduce its length.

/125...

The summary of the analysis of the total group is given in (2) where it can be seen that all the predictors are highly significant except "Year of entry" which is not significant considering that there are six other predictors (the Bonferroni inequality implying significance only at the 7x2.4% = 17% level). Note that the first predictor, of whether or not the student qualified, is significant at this stage of the analysis, lending support to the prejudice against some matriculation boards. Once other factors have been taken into account, however, this predictor is no longer significant, as the subsequent analysis reveals.

The cross tabulation of the dependent variable with each predictor is given in (12) and (13), and the tabulation with the "best" grouping of the categories of the predictor is given in (14) and (15). The detailed analysis of each predictor is given in (8) to (11) and shows how the final grouping of categories within a predictor was constructed, together with competing possible merges and splits at each stage. This detail is necessarily voluminous and in most of the analyses only portions are of primary interest to the user. Note for example that in (10) the missing information category ('?') in the mathematics mark predictor is merged with a 'D' symbol in the very first stage of the analysis. Thus the four students who have missing information on this predictor (see (13)) may be recoded to a 'D' symbol for a non-CHAID analysis.

/126...

"Number of commercial subjects" is the most significant predictor, and the table on (2) (which is a copy of the same table in (14)) shows clearly that NSC students tend to take more commercial subjects (generally considered "soft" options) as compared with the JMB and TSSC & TUEC students. The total group is now subdivided into four subgroups according to the specifications on (2). Because the options for this analysis specified that subgroups with less than one hundred students should not be analysed further (a high setting to reduce the volume of output for this discussion), those with two or more commercial subjects are not considered again.

Page (3) summarizes the analysis of those students who did not offer any commercial subjects in their matriculation. The first four predictors are now all found to be not significant, as well as the last predictor "English mark" which is significant when considered in isolation, but not in conjunction with the six other predictors (the Bonferroni inequality implying significance only at the 7x2.07% = 14% level). The most significant predictor is "Type of school" and the table on (3) shows that no TSSC & TUEC or "Other" students attend the so-called "cram" colleges.

The analysis of those students who offered a single commercial subject in the matriculation is summarized in (4). Here again the first four predictors are found to be not significant. As above, the most significant predictor is "Type of school" and the TSSC & TUEC and "Other" students are again underrepresented in the "cram" colleges.

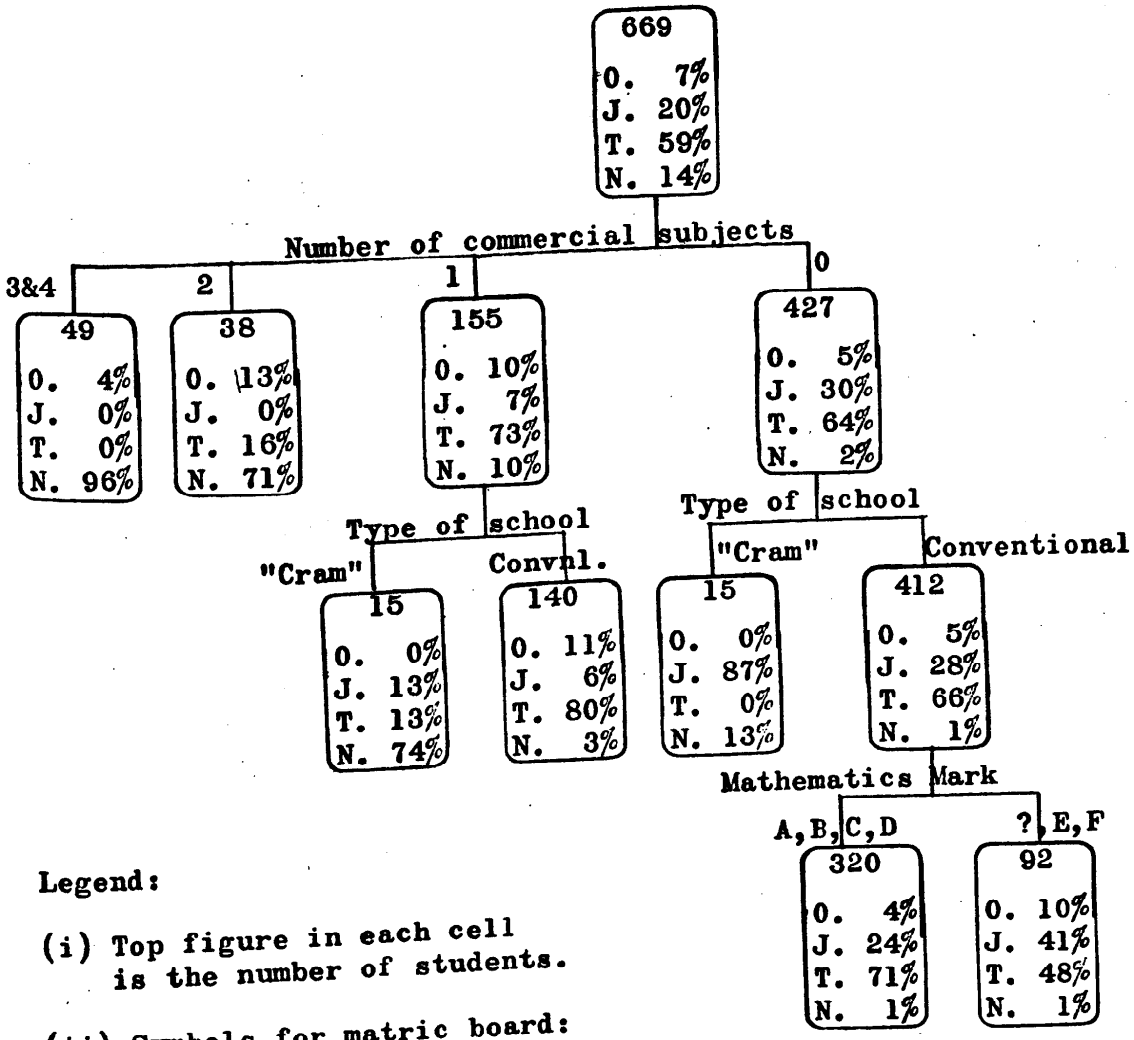
It may be noted that only a very small minority of students who offer one or no commercial subjects for their matriculation attend "cram" colleges. The majority of the 99 students from "cram" colleges (see last table on (12)) have already been separated from the analysis on the basis of having two or more commercial subjects in their matriculation curriculum.

Finally, (5) summarizes the analysis of those at conventional schools that included no commercial subjects in their matriculation. The only predictor of any significance is "mathematics mark" which shows that the JMB students seem to score worse than those under other matriculation boards - perhaps due to stiffer examinations. Pages (6) and (7) conclude the analysis showing that there are no more significant divisions. Stringent options have been selected for this analysis in order to curtail the output - in practice "looser" options are chosen, to provide further information.

For simplicity of presentation the above results are summarized in the dendrogram of figure 9.1. (see next page).

/128...

Figure 9.1. Dendrogram of the analysis of matriculation boards.



Legend:

(i) Top figure in each cell is the number of students.

(ii) Symbols for matric board:

0. = Other
J. = JMB
T. = TSSC & TUEC
N. = NSC

/129...

CHAPTER 10. Summary and Conclusions.

10.1 Summary.

A number of extensions to A.I.D. have been presented and discussed. An implementation of these ideas has been provided to analyse nominal dependent variables, though of course ordinal dependent variables could be categorized and reduced to nominal. The major extension of the technique is the utilization of the experience gained in the first part of this thesis concerning the inferential aspects of A.I.D. In this second part, the "best" split is defined as the "most significant" split rather than the "most explanatory" which does not take into account the type of the predictor nor the number of its categories. This introduction of continual significance testing at each stage of the analysis has provided a criterion for assessing multiway subdivisions of the data, which is felt to yield a more effective analysis than the traditional binary splits that are occasionally misleading.

The problem of predictors that are not monotonic (ordinal) nor yet fully free (nominal) has been considered briefly. The special case of a predictor that is monotonic except for a single category that usually denotes missing or unobtainable information, has resulted in a floating predictor that has been found to be of great practical value.

/130...

Like its predecessor A.I.D., CHAID is a powerful technique for partitioning data into more homogeneous groups. While this type of analysis is often used as a precursor to a more parametric technique, it has been found frequently to be an end analysis in itself. This is especially true in those fields where the typical method of analysis has been to produce and examine (if it is humanly possible) all cross tabulations of the data. CHAID automates this to some extent by rejecting insignificant cross tabulations and immediately focussing the researcher's attention on the potentially useful subdivisions.

10.2 Conclusions.

There is much room for improvement and added sophistication to A.I.D. type techniques, and to CHAID in particular. Even while working on CHAID, and especially during live data analysis many extensions and ideas for future research came to mind. We briefly consider some of these ideas in this conclusion.

Firstly, note that the dependent variable has been considered as a purely nominal (ie. free) variable. If it takes measures on an interval or ratio scale then standard A.I.D. is the applicable technique. Ordinal dependent variables (ie. monotonic) or modifications thereto (eg. a floating or interleaved-monotonic variable) have not been considered here. The problem of contingency tables with ordered categories has received

/131...

attention in the literature and the test statistics developed for various models (see for example Simon (1974) who also briefly reviews other work in this field) could be used as an appropriate criterion in analysing data with a monotonic dependent variable. A second approach to this problem which could be applied equally well to a floating dependent variable as well as other types is to transpose the CHAID analysis and condense also the dependent variable by merging its categories. This would result in an even more parsimonious description of the data. Perhaps ideally both the above approaches could be effected simultaneously.

The second major improvement concerns the particular implementation of CHAID that has been presented which relies on moderate to large samples in order that the chi-square test for contingency tables may be applicable. As mentioned earlier, an exact test could be used for small samples though because of the large number of contingency tables to be examined this could prove cumbersome despite the speed of present day computers. A possible alternative, especially useful when particular rows or columns have few observations (possibly a result of the CHAID splitting process), is to force such rows or columns to be merged, a process that is often performed in practice when insufficient numbers of observations occur in some "cell" in a chi-square test. More in line with the philosophy of this thesis would be to follow the literature on contingency tables with small expected values, for example

/132...

Nass (1959) for large sparse tables, or Margolin and Light (1974) who examine small sample behaviour in two-way contingency tables for a variety of test statistics. These alternatives for small samples would need to be studied more closely before their implementation in a CHAID-type program, especially to ensure that the procedure would still converge to a stable solution.

Finally, we reiterate that computer analyses are no substitute for sound commonsense and the researcher's own extensive experience in his field. In order that the user may impose his knowledge and experience on the progress of the analysis, user specifications of primary subdivisions of the data (as permitted in AID-III) is a desirable addition to our implementation of CHAID. In this way the researcher can control the analysis guided by the computer - a combination that we hope will be profitable in enhancing the understanding of real world phenomena.

/133...

LIST OF REFERENCES.

ANDERSON,T.W.(1971). The Statistical Analysis of Time Series. New York:Wiley.

ASSAEL,H.(1970). Segmenting markets by group purchasing behaviour: An application of the A.I.D. technique. J. Marktng Res., 7, 153-158.

BELSON,W.A.(1959). Matching and prediction on the principle of biological classification. Appl. Statist., 8, 65-75.

BRADLEY,J.V.(1968). Distribution Free Methods. Englewood Cliffs, New Jersey:Prentice-Hall.

BURR,R.W.(1967). A useful approximation to the normal distribution function, with application to simulation. Technometrics, 9, 647-651.

COCHRAN,W.G.(1963). Sampling Techniques. New York:Wiley.

CONOVER,W.J.(1971). Practical Nonparametric Statistics. New York:Wiley.

CRAMER,E.M.(1971). Book Review. Psychometrika, 36, 440-442.

CURNOW,R.N. and DUNNETT,C.W.(1962). The numerical evaluation of certain multivariate normal integrals. Ann. Math. Statist., 33, 571-579.

DOYLE,P.(1973). The use of automatic interaction detector and similar search procedures. Operat. Res. Quart., 24, 465-467.

EFROYMSON,M.A.(1965). Multiple regression analysis. In Ralston and Wilf: Mathematical methods for digital computers. New York:Wiley.

FELLER,W.(1968). An Introduction to Probability Theory and its Applications. New York:Wiley.

GILLO,M.W.(1972). MAID, a Honeywell 600 program for an automatic survey analysis. Behavioral Science, 17, 251-252.

GILLO,M.W. and SHELLY,M.W.(1974). Predictive modeling in multivariable and multivariate data. J. Amer. Statist. Ass., 69, 646-653.

GRIZZLE,J.E.(1967). Continuity correction in the χ^2-test for 2 x 2 tables. The Amer. Statist., 21, 28-32.

GUPTA,S.S.(1963). Probability integrals of multivariate normal and multivariate t. Ann. Math. Statist., 34, 792-837.

HALD,A.(1952). Statistical Tables and Formulas. New York:Wiley.

/134...

HAMMERSLEY,J.M. and HANDSCOMB,D.C.(1964). Monte Carlo Methods. London:Methuen.

HASTINGS,C.(1955). Approximations for Digital Computers. Princeton University Press.

HAWKINS,D.M.(1975). Testing a sequence of observations for a shift in location. Submitted to J. Amer. Statist. Ass.

HAWKINS,D.M. and MERRIAM,D.F.(1973). Optimal zonation of digitized sequential data. Math. Geology, 5, 389-395.

HEALD,G.I.(1972). The application of the A.I.D. programme and multiple regression techniques to the assessment of store performance and site selection. Operat. Res. Quart., 23, 445-457.

HEISE,D.R.(1971). Book Review. Public Opinion Quart., 35, 307-309.

HILL,M.O.(1974). Correspondence analysis: a neglected multivariate method. Appl. Statist., 23, 340-354.

HUNT,E.B.,MARIN,J. and STONE,P.J.(1966). Experiments in Induction. New York:Academic Press.

IBM(1968). System 360 Operating System. PL/1 Subroutine Library. Computational Subroutines. Form C28-6590. Hursly Park, England.

IBM(1970). Scientific Subroutine Package, Version III, Programmer's Manual. Form GH20-0205. New York.

KENDALL,M.G. and STUART,A.(1958). The Advanced Theory of Statistics, Vol. 1. London:Griffin.

_____(1961). The Advanced Theory of Statistics, Vol. 2. London:Griffin.

KNUTH,D.E.(1969). The Art of Computer Programming, Vol. 2, Seminumerical Algorithms. Reading, Mass.:Addison-Wesley.

KRISHNAIAH,P.R. and ARMITAGE,J.V.(1965). Tables for the distribution of correlated chi-square variates with one degree of freedom. ARL 65-136, Aerospace Research Laboratories, Wright-Patterson Air Force Base, Ohio.

LAWLEY,D.N. and MAXWELL,A.E.(1971). Factor Analysis as a Statistical Method, 2nd ed. London:Butterworth.

MACNAUGHTON-SMITH,P.(1963). The classification of individuals by the possession of attributes with a criterion. Biometrics, 19, 364-366.

/135...

MARGOLIN,B.H. and LIGHT,R.J.(1974). An analysis of variance for categorical data II: Small sample comparisons with chi-square and other competitors. J. Amer. Statist. Ass., 69, 755-764.

MARSAGLIA,G.,MACLAREN,M.D. and BRAY,T.A.(1964). A fast procedure for generating normal random variables. Comm. Assoc. Comp. Mach., 7, 4-10.

MAXWELL,A.E.(1961). Analysing Qualitative Data. London:Methuen.

MCGEE,V.E. and CARLETON,W.T.(1970). Piecewise regression. J. Amer. Statist. Ass., 65, 1109-1124.

MESSENGER,R.C. and MANDELL,L.M.(1972). A modal search technique for predictive nominal scale multivariate analysis. J. Amer. Statist. Ass., 67, 768-772.

MEYER,D.L.(1969). Methods of generating random normal numbers. Educ. and Psycho. Measurement, 29, 193-198.

MILLER,R.G.(1966). Simultaneous Statistical Inference. New York:McGraw-Hill.

MILTON,R.C.(1972). Computer evaluation of the multivariate normal integral. Technometrics, 14, 881-889.

MORGAN,J.N. and MESSENGER,R.C.(1973). THAID - A sequential analysis program for the analysis of nominal scale dependent variables. Michigan:Survey Research Centre, Institute for Social Research, University of Michigan.

MORGAN,J.A.,SIRAGELDIN,I. and BAERWALDT,N.(1965). Productive Americans. Michigan: Institute for Social Research, University of Michigan.

MORGAN,J.A. and SONQUIST,J.N.(1963a). Problems in the analysis of survey data: and a proposal. J. Amer. Statist. Ass., 58, 415-434.

————(1963b). Some results from a non-symmetrical branching process that looks for interaction effects. Proc. of the Soc. Stats. Sec., ASA, 40-53.

NASS,C.A.G.(1959). The χ^2 test for small expectations in contingency tables, with special reference to accidents and absenteeism. Biometrika, 46, 365-385.

ORR,L.(1972). The dependence of transition proportions in the education system on observed social factors and school characteristics. J. R. Statist. Soc. A, 135, 74-95.

/136...

MARGOLIN,B.H. and LIGHT,R.J.(1974). An analysis of variance for categorical data II: Small sample comparisons with chi-square and other competitors. J. Amer. Statist. Ass., 69, 755-764.

MARSAGLIA,G.,MACLAREN,M.D. and BRAY,T.A.(1964). A fast procedure for generating normal random variables. Comm. Assoc. Comp. Mach., 7, 4-10.

MAXWELL,A.E.(1961). Analysing Qualitative Data. London:Methuen.

MCGEE,V.E. and CARLETON,W.T.(1970). Piecewise regression. J. Amer. Statist. Ass., 65, 1109-1124.

MESSENGER,R.C. and MANDELL,L.M.(1972). A modal search technique for predictive nominal scale multivariate analysis. J. Amer. Statist. Ass., 67, 768-772.

MEYER,D.L.(1969). Methods of generating random normal numbers. Educ. and Psycho. Measurement, 29, 193-198.

MILLER,R.G.(1966). Simultaneous Statistical Inference. New York:McGraw-Hill.

MILTON,R.C.(1972). Computer evaluation of the multivariate normal integral. Technometrics, 14, 881-889.

MORGAN,J.N. and MESSENGER,R.C.(1973). THAID - A sequential analysis program for the analysis of nominal scale dependent variables. Michigan:Survey Research Centre, Institute for Social Research, University of Michigan.

MORGAN,J.A.,SIRAGELDIN,I. and BAERWALDT,N.(1965). Productive Americans. Michigan: Institute for Social Research, University of Michigan.

MORGAN,J.A. and SONQUIST,J.N.(1963a). Problems in the analysis of survey data: and a proposal. J. Amer. Statist. Ass., 58, 415-434.

_____(1963b). Some results from a non-symmetrical branching process that looks for interaction effects. Proc. of the Soc. Stats. Sec., ASA, 40-53.

NASS,C.A.G.(1959). The χ^2 test for small expectations in contingency tables, with special reference to accidents and absenteeism. Biometrika, 46, 365-385.

ORR,L.(1972). The dependence of transition proportions in the education system on observed social factors and school characteristics. J. R. Statist. Soc. A, 135, 74-95.

/136...

OWEN,D.(1962). Handbook of Statistical Tables. Reading, Mass.: Addison-Wesley.

PREECE,D.A.(1971). Iterative procedures for missing values in experiments. Technometrics, 13, 743-753.

PRESS,L.I.,ROGERS,M.S. and SHURE,G.H.(1969). An interactive technique for the analysis of multivariate data. Behavioral Science, 14, 364-370.

PURI,M.L. and SEN,P.K.(1971). Nonparametric Methods in Multivariate Analysis. New York:Wiley.

ROSS,J.A. and BANG,S.(1966). The A.I.D. computer program, used to predict adoption of family planning in Koyang. Popul. Studies, 20, 61-75.

SIMON,G.(1974). Alternative analyses for the singly-ordered contingency table. J. Amer. Statist. Ass., 69, 971-976.

SONQUIST,J.N.(1970). Multivariate Model Building. Michigan: Institute for Social Research, University of Michigan.

SONQUIST,J.N.,BAKER,E.L. and MORGAN,J.A.(1971). Searching for Structure (Alias-AID-III). Michigan: Institute for Social Research, University of Michigan.

SONQUIST,J.N. and MORGAN,J.A.(1964). The Detection of Interaction Effects. Monograph No. 35. Survey Research Centre, Institute for Social Research, University of Michigan.

STONE,M.(1974). Cross-validatory choice and assessment of statistical predictions. J. R. Statist. Soc. B, 36, 111-147.

TAUSWORTHE,R.C.(1965). Random numbers generated by linear recurrence modulo two. Math. Comput., 19, 201-209.

WHITTLESEY,J.R.B.(1968). A comparison of the correlational behavior of random number generators for the IBM 360. Comm. of the ACM, 11, 641-644.

YEARSLEY,R.B. and GRAHAM,G.M.R.(1973). Handbook of Computer Management. Epping:Gower Press.

Subroutine NORSP

Subroutine NORSP is a special purpose subroutine written in IBM S/360/370 Assembler language to return a vector of independent random normal deviates, with possible restriction of range on any set of deviates. It is invoked from Fortran by

CALL NORSP(N,ISEED,ITST,P,PR)

where

N Input fullword integer. The number of random normal deviates required

ISEED Input/Output fullword. A seed for the pseudo-random number generatorn

Input: First bit must be 0. At least one bit must be 1.

Output: New seed that can be used in a subsequent call to NORSP.

ITST Input fullword integer array of length N.
Indicates the range of the corresponding normal deviate.

= 0 No restriction on range.

> 0 Used for range of type $|x|>a$. See PR below.

< 0 Used for range of type $|x|<a$. See PR below.

P Output fullword real array of length N.
Contains the random normal deviates.

/A 2 ....

PR Input fullword real array of length 3. Used in conjunction with ITST to further define the range of the normal deviate. These constants modify the value of the Uniform random number U from which the random normal deviate is obtained by the inverse transformation.

If ITST>0 then new Uniform number is $UxPR(3) + \delta PR(1)$

where $\delta = 0$ if $UxPR(3) < PR(2)$

$= 1$ otherwise.

and if ITST<0 then new Uniform number is

$UxPR(1) + PR(2)$

Notes on the constants PR.

If the required range is $|x| \gtrless a$ then

$PR(1) = 2\Phi(a) - 1$

$PR(2) = 1 - \Phi(a) = \frac{1}{2}(1 - PR(1))$

$PR(3) = 2 - 2\Phi(a) = 2xPR(2)$

where $\Phi(a)$ is the cumulative normal distribution function.

The routine could also be used to generate random normal deviates within other ranges, for example if the required range is $a < x < b$ then

$PR(1) = \Phi(b) - \Phi(a)$

$PR(2) = \Phi(a)$ and use ITST > 0

while if the range is $x < a$ or $x > b$ then also

$PR(3) = 1 - \Phi(b) + \Phi(a)$ and use ITST < 0

Finally, note that the above holds even if $a = -\infty$ or $b = \infty$.

/A 3 ....

The Assembler listing of the routine follows. The routine requires tables of the inverse cumulative normal distribution IVT1 and IVT2 as described in the text. A listing of these tables follows the Assembler routine. Fortran subroutine HAST(U,X) which is required to convert the probability U ($<10^{-4}$) to the normal deviate X, follows the tables.

LISTING OF SUBROUTINE NORSP

```
NORSP   START   0               (N,ISEED,ITST,P,PR(3))
        USING   *,10            USE R10 AS BASE ADDRESS
        SAVE    (14,12),T,*     SAVE ALL REGISTERS
        LR      10,15           LOAD ENTRY POINT INTO R10
        ST      13,R13          SAVE R13
        LA      13,SA           POINT TO NEW S.A. IN CASE WE USE IT
        L       2,0(0,1)        ADDR OF NO. OF ITS
        L       2,0(0,2)        NO. OF ITS
        L       11,4(0,1)       ADDR OF SEED
        L       7,0(0,11)       SEED
        L       3,8(0,1)        ADDR OF RANGE-TYPE
        L       4,12(0,1)       ADDR OF ANSWERS
        L       5,16(0,1)       ADDR OF REC CONSTANTS
        SR      6,6             INCREMENT OF RANGE-TYPE & ANSWERS
LOOP    SDR     0,0             OBTAIN RANDOM NUMBER (R.N.) 0 - 1
        ST      7,WORKA+4       SAVE SOME RANDOM BITS AT THE END
        LR      8,7             COPY TO R8
        SRL     8,13            RIGHT SHIFT
        XR      8,7             EXCL OR
        LR      7,8             COPY BACK
        SLL     8,18            LEFT SHIFT
        XR      7,8             EXCL. OR
        SLL     7,1             REMOVE SIGN BIT
        ST      7,WORKB         RANDOM BITS IN WORKB
        SRL     7,1             RESET FOR NEW SEED
        MVC     WORKA+1(4),WORKB    R.N. IN WORKA
        AD      0,WORKA         OUR NORMALISED R.N.
        SR      8,8             FOR COMPARISON
        C       8,0(6,3)        FIND HOW TO MODIFY THE R.N.
        BE      NORDEV          NO MODIFICATION REQ.
        BH      INSID           INNER RANGE
        ME      0,8(0,5)        RE-ADJUST RANGE OF R.N.
        CE      0,4(0,5)        IN BOTTOM OR TOP PART OF RANGE?
        BNH     NORDEV          IN BOTTOM PART ALREADY
        AE      0,0(0,5)        SHIFT TO TOP PART
        B       NORDEV          GET ON WITH IT.
INSID   ME      0,0(0,5)        RESCALE R.N.
        AE      0,4(0,5)        SHIFT TO MIDDLE OF RANGE
NORDEV  SR      9,9             SET INDICATOR (R.N. < .5)
        CE      0,=E'0.5'       R.N. < 0.5 ?
        BNH     N2              IT IS LESS
        LA      9,1             RESET INDICATOR (R.N. > .5)
        SE      0,ONE           SUBTRACT ONE.
        LPER    0,0             GIVING (1.0 - R.N.)
```

LISTING OF SUBROUTINE NORSP CONTINUED

```
N2      CE      0,=E'0.01'      CAN WE USE TABLE 1 ?
        BNL     LOOK1           YES WE CAN
        CE      0,=E'1.0E-4'      CAN WE USE TABLE 2  ?
        BNL     LOOK2           YES WE CAN
        STE     0,RV            STORE R.N. IN CORE
        CALL    FAST,(RV,ANS)     USE HASTING'S APPROXIMATION
        LE      4,ANS           ANSWER IN FL. R. 4
        B       ENDOFF          CARRY ON AS USUAL
LOOK2   LE      2,=E'1.0E5'     APPLICABLE CONSTANT
        L       12,TAB2         ADDR OF APPLICABLE TABLE
        B       FIND            NOW DO TABLE LOOK-UP ETC.
LOOK1   LE      2,=E'2.0E3'     APPLICABLE CONSTANT FOR PRODUCING INDEX
        L       12,TAB1         ADDR OF APPLICABLE TABLE
FIND    MER     2,0             FOR ELEMENT OF TABLE
        LD      4,MASK          TO CONVERT TO INTEGER.
        AUR     4,2             GET THE INTEGER PART
        STE     4,ANS           TO GET IT INTO A FIXED REGISTER
        L       14,ANS          PUT INTO R14
        N       14,=X'00FFFFFF'    REMOVE EXPONENT
        SLL     14,2            FOR ELEMENT IN TABLE.
        SDR     2,4             FRACTION ABOVE ELEMENT IN TABLE.
        LE      4,4(14,12)      TAB(K+1)
        SE      4,0(14,12)      TAB(K+1) - TAB(K)
        MER     4,2             REM*(T(K+1) - T(K))
        AE      4,0(14,12)      T(K) + REM*(T(K+1) - T(K))
ENDOFF  LTR     9,9             MODIFY ANSWER?
        BZ      ANSWER          NO
        LNER    4,4             X = -X
ANSWER  STE     4,0(6,4)        STORE ANSWER
        LA      6,4(6)          ADD 4 TO R6
        BCT     2,LOOP          DO APPROPRIATE NO. OF TIMES
        ST      7,0(0,11)       RESTORE NEW SEED
        L       13,R13          LAST PROGRAM'S S.A. ADDRESS
        RETURN (14,12),T,RC=0   GO BACK TO CALLING PROGRAM
        DS      0D              ALIGN ON DOUBLE WORD
WORKA   DC      X'4CCCCCCCCCCCCCCC'
WORKB   DC      X'4CCCCCCCCCCCCCCC'
MASK    DC      X'460CCCCCCCCCCCCC'   AN APPROPRIATE ZERO.
TAB1    DC      V(IVT1)         ADDRESS OF FIRST TABLE
TAB2    DC      V(IVT2)         ADDRESS OF SECOND TABLE
SA      DS      18F             REGISTER SAVE AREA
R13     DC      F'0'            SAVE AREA FOR R13
ONE     DC      E'1.0'          FLOATING POINT 1
RV      DC      E'0.0'          INPUT TO SUBROUTINE HAST
ANS     DC      E'0.0'          OUTPUT FROM SUBROUTINE HAST
        END     NORSP
```

Listing of Table IVT1 as required by subroutine NORSP.

(Note that the first two lines of the table are not used by NORSP)

	.0000	.0005	.0010	.0015	.0020	.0025	.0030	.0035	.0040	.0045
	.0050	.0055	.0060	.0065	.0070	.0075	.0080	.0085	.0090	.0095
0.0	6.581053	3.286602	3.084897	2.965596	2.876983	2.806281	2.747256	2.696456	2.651770	2.611816
	2.575634	2.542538	2.512008	2.483651	2.457162	2.432290	2.408836	2.386637	2.365555	2.345474
0.01	2.326296	2.307938	2.290324	2.273396	2.257092	2.241369	2.226180	2.211489	2.197259	2.183461
	2.170066	2.157050	2.144389	2.132063	2.120052	2.108340	2.096910	2.085748	2.074839	2.064172
0.02	2.053735	2.043516	2.033507	2.023697	2.014079	2.004643	1.995382	1.986289	1.977359	1.968582
	1.959955	1.951470	1.943125	1.934913	1.926828	1.918868	1.911028	1.903303	1.895691	1.888186
0.03	1.880787	1.873488	1.866289	1.859185	1.852174	1.845252	1.838418	1.831669	1.825002	1.818415
	1.811906	1.805472	1.799113	1.792826	1.786608	1.780459	1.774377	1.768360	1.762406	1.756514
0.04	1.750682	1.744909	1.739193	1.733534	1.727930	1.722380	1.716883	1.711436	1.706040	1.700693
	1.695395	1.690143	1.684937	1.679777	1.674662	1.669589	1.664559	1.659572	1.654625	1.649717
0.05	1.644851	1.640021	1.635232	1.630477	1.625761	1.621079	1.616434	1.611823	1.607244	1.602702
	1.598190	1.593712	1.589265	1.584849	1.580464	1.576110	1.571784	1.567489	1.563221	1.558982
0.06	1.554771	1.550588	1.546431	1.542300	1.538197	1.534119	1.530066	1.526037	1.522034	1.518055
	1.514099	1.510168	1.506259	1.502375	1.498511	1.494670	1.490851	1.487054	1.483278	1.479524
0.07	1.475790	1.472075	1.468383	1.464708	1.461055	1.457420	1.453805	1.450209	1.446630	1.443071
	1.439530	1.436007	1.432501	1.429013	1.425543	1.422089	1.418653	1.415233	1.411829	1.408442
0.08	1.405070	1.401715	1.398376	1.395051	1.391743	1.388449	1.385170	1.381907	1.378657	1.375422
	1.372203	1.368997	1.365804	1.362626	1.359462	1.356310	1.353172	1.350048	1.346937	1.343839
0.09	1.340754	1.337681	1.334620	1.331573	1.328538	1.325515	1.322504	1.319505	1.316517	1.313541
	1.310577	1.307625	1.304684	1.301754	1.298835	1.295927	1.293031	1.290145	1.287270	1.284405
0.10	1.281550	1.278707	1.275873	1.273049	1.270236	1.267433	1.264640	1.261856	1.259083	1.256319
	1.253564	1.250819	1.248083	1.245358	1.242640	1.239932	1.237233	1.234544	1.231862	1.229190
0.11	1.226527	1.223872	1.221226	1.218588	1.215960	1.213338	1.210727	1.208121	1.205525	1.202937
	1.200358	1.197786	1.195221	1.192665	1.190117	1.187576	1.185042	1.182517	1.179999	1.177489
0.12	1.174986	1.172490	1.170001	1.167520	1.165046	1.162579	1.160118	1.157666	1.155220	1.152781
	1.150349	1.147923	1.145504	1.143091	1.140686	1.138288	1.135895	1.133510	1.131130	1.128757
0.13	1.126390	1.124029	1.121675	1.119328	1.116985	1.114651	1.112320	1.109997	1.107679	1.105368
	1.103062	1.100762	1.098467	1.096179	1.093896	1.091619	1.089349	1.087082	1.084822	1.082568
0.14	1.080318	1.078074	1.075836	1.073604	1.071377	1.069154	1.066936	1.064725	1.062518	1.060317
	1.058121	1.055929	1.053743	1.051562	1.049387	1.047215	1.045050	1.042888	1.040731	1.038579
0.15	1.036432	1.034290	1.032153	1.030021	1.027893	1.025769	1.023651	1.021536	1.019426	1.017323
	1.015222	1.013125	1.011034	1.008946	1.006864	1.004786	1.002710	1.000641	0.998576	0.996515

	.CCC0	.CCC5	.C01C	.CC15	.C020	.0C25	.0030	.0035	.C04C	.0045
	.0050	.0055	.0C60	.C065	.007C	.0075	.0080	.0085	.0090	.0095
0.16	0.994458	C.992405	C.990356	C.988311	C.986271	0.984235	0.982202	0.980174	0.978150	0.976130
	0.974114	0.972101	0.970093	0.968089	0.966088	0.964091	0.962098	0.960109	0.958124	0.956143
0.17	C.954165	0.952191	C.950221	C.948254	0.946291	0.944332	0.942376	0.940424	C.938475	0.936530
	C.934589	C.932651	C.930717	C.928786	0.926858	0.924934	0.923014	0.921096	0.919183	0.917272
0.18	0.915365	C.913461	0.911560	0.909663	0.907769	C.905879	0.903991	0.902107	0.900226	0.898348
	0.896473	C.894602	0.892733	C.890868	C.889006	0.887146	0.885290	0.883437	C.881587	0.879740
0.19	0.877896	C.876055	0.874217	0.872382	0.870550	0.868720	0.866894	0.865070	0.863250	0.861432
	0.859617	C.857805	0.855996	C.854189	0.852386	0.850585	0.848786	0.846991	C.845198	0.843409
0.20	0.841621	C.839836	C.838054	C.836275	0.834499	C.832725	0.830953	0.829184	C.827418	C.825655
	C.823893	C.822135	C.820379	0.818626	0.816875	0.815126	0.813380	0.811637	0.809896	C.808157
0.21	C.806421	C.804687	0.802956	C.801227	0.799501	C.797777	0.796055	0.794336	C.792619	0.790904
	0.789191	0.787481	0.785774	C.784068	0.782365	0.780664	0.778965	0.777269	C.775575	C.773883
0.22	0.772193	0.770505	C.768820	C.767137	0.765456	C.763777	0.762100	0.760426	0.758753	0.757083
	0.755415	C.753749	C.752085	C.750423	0.748763	0.747105	0.745449	0.743796	0.742144	0.740494
0.23	C.738847	C.737201	C.735557	C.733916	C.732276	C.730638	0.729003	0.727369	0.725737	C.724107
	0.722479	0.720853	0.719229	0.717606	0.715986	0.714367	0.712751	0.711136	0.709523	0.707912
0.24	C.706302	C.704695	0.703089	C.701486	C.699883	0.698283	0.696685	0.695088	C.693493	C.691900
	0.690309	0.688719	C.687131	C.685545	0.683961	0.682378	0.680797	0.679217	0.677640	0.676064
0.25	C.674490	C.672917	0.671346	C.669777	0.668209	0.666643	0.665079	0.663516	C.661955	0.660395
	0.658838	0.657291	0.655727	0.654173	0.652622	0.651072	0.649523	0.647977	0.646431	0.644888
0.26	0.643345	C.641805	C.640265	0.638728	C.637192	C.635657	0.634124	0.632592	C.631062	0.629533
	0.628006	0.626480	0.624956	0.623433	0.621911	0.620391	0.618873	0.617356	0.615840	0.614326
0.27	C.612813	C.611301	C.609791	0.608283	0.606775	0.605269	0.603765	0.602261	C.600760	0.599259
	0.597760	C.596262	0.594766	0.593271	0.591777	0.590284	0.588793	0.587303	0.585815	0.584327
0.28	0.582841	0.581357	0.579873	C.578391	0.576910	0.575431	0.573952	0.572475	0.570999	0.569525
	0.568051	0.566579	C.565108	C.563639	0.562170	0.560703	0.559237	0.557772	0.556308	0.554846
0.29	0.553385	0.551925	0.550466	C.549008	0.547551	0.546096	0.544642	0.543188	0.541736	C.540286
	0.538836	C.537387	0.535940	C.534494	0.533048	0.531604	0.530161	0.528719	C.527279	0.525839
0.30	0.524400	C.522963	0.521527	C.520091	0.518657	C.517224	0.515791	0.514360	0.512930	0.511501
	0.510073	0.508646	0.507221	C.505796	C.504372	0.502949	0.501527	0.500107	0.498687	0.497268
0.31	0.495850	0.494434	0.493018	C.491603	0.490189	0.488776	0.487364	0.485954	0.484544	0.483135
	0.481727	C.480320	0.478914	0.477509	0.476104	0.474701	0.473299	0.471897	C.470497	0.469097
0.32	0.467699	0.466301	0.464904	0.463508	0.462113	0.460719	0.459326	0.457934	0.456542	0.455152
	0.453762	C.452373	0.450985	C.449598	C.448212	0.446827	0.445442	0.444059	C.442676	0.441294

	.0000	.0005	.0010	.0015	.0020	.0025	.0030	.0035	.0040	.0045
	.0050	.0055	.0060	.0065	.0070	.0075	.0080	.0085	.0090	.0095
0.33	0.439913	0.438533	0.437153	0.435775	0.434397	0.433020	0.431644	0.430269	0.428894	0.427521
	0.426148	0.424776	0.423405	0.422034	0.420665	0.419296	0.417928	0.416560	0.415194	0.413828
0.34	0.412463	0.411099	0.409735	0.408373	0.407011	0.405650	0.404289	0.402930	0.401571	0.400212
	0.398855	0.397498	0.396142	0.394787	0.393432	0.392079	0.390726	0.389373	0.388022	0.386671
0.35	0.385320	0.383971	0.382622	0.381274	0.379926	0.378580	0.377234	0.375888	0.374543	0.373199
	0.371856	0.370513	0.369171	0.367830	0.366489	0.365149	0.363810	0.362471	0.361133	0.359796
0.36	0.358459	0.357123	0.355787	0.354452	0.353118	0.351784	0.350451	0.349119	0.347787	0.346456
	0.345125	0.343796	0.342466	0.341138	0.339809	0.338482	0.337155	0.335829	0.334503	0.333178
0.37	0.331853	0.330529	0.329206	0.327883	0.326561	0.325239	0.323918	0.322598	0.321278	0.319958
	0.318639	0.317321	0.316003	0.314686	0.313369	0.312053	0.310738	0.309423	0.308108	0.306794
0.38	0.305481	0.304168	0.302855	0.301544	0.300232	0.298921	0.297611	0.296301	0.294992	0.293683
	0.292375	0.291067	0.289760	0.288453	0.287147	0.285841	0.284535	0.283231	0.281926	0.280622
0.39	0.279319	0.278016	0.276714	0.275412	0.274110	0.272809	0.271508	0.270208	0.268909	0.267609
	0.266311	0.265012	0.263714	0.262417	0.261120	0.259823	0.258527	0.257232	0.255936	0.254641
0.40	0.253347	0.252053	0.250759	0.249466	0.248174	0.246881	0.245589	0.244298	0.243007	0.241716
	0.240426	0.239136	0.237847	0.236558	0.235269	0.233981	0.232693	0.231405	0.230118	0.228831
0.41	0.227545	0.226259	0.224973	0.223688	0.222403	0.221119	0.219835	0.218551	0.217267	0.215984
	0.214701	0.213419	0.212137	0.210855	0.209574	0.208293	0.207013	0.205732	0.204452	0.203173
0.42	0.201893	0.200614	0.199336	0.198058	0.196780	0.195502	0.194225	0.192948	0.191671	0.190394
	0.189118	0.187843	0.186567	0.185292	0.184017	0.182743	0.181468	0.180194	0.178921	0.177647
0.43	0.176374	0.175101	0.173829	0.172557	0.171285	0.170013	0.168741	0.167470	0.166199	0.164929
	0.163658	0.162388	0.161119	0.159849	0.158580	0.157311	0.156042	0.154773	0.153505	0.152237
0.44	0.150969	0.149702	0.148434	0.147167	0.145900	0.144634	0.143367	0.142101	0.140835	0.139570
	0.138304	0.137039	0.135774	0.134509	0.133245	0.131980	0.130716	0.129452	0.128188	0.126925
0.45	0.125661	0.124398	0.123135	0.121872	0.120610	0.119348	0.118085	0.116823	0.115562	0.114300
	0.113039	0.111777	0.110516	0.109255	0.107995	0.106734	0.105474	0.104213	0.102953	0.101693
0.46	0.100434	0.099174	0.097915	0.096655	0.095396	0.094137	0.092879	0.091620	0.090361	0.089103
	0.087845	0.086587	0.085329	0.084071	0.082813	0.081556	0.080298	0.079041	0.077784	0.076527
0.47	0.075270	0.074013	0.072756	0.071500	0.070243	0.068987	0.067731	0.066475	0.065219	0.063963
	0.062707	0.061451	0.060195	0.058940	0.057684	0.056429	0.055174	0.053919	0.052664	0.051409
0.48	0.050154	0.048899	0.047644	0.046389	0.045135	0.043880	0.042626	0.041371	0.040117	0.038863
	0.037608	0.036354	0.035100	0.033846	0.032592	0.031338	0.030084	0.028830	0.027576	0.026323
0.49	0.025069	0.023815	0.022562	0.021308	0.020054	0.018801	0.017547	0.016294	0.015040	0.013787
	0.012533	0.011280	0.010027	0.008773	0.007520	0.006267	0.005013	0.003760	0.002507	0.001253
0.50	-0.000000									

Listing of Table IVT2 as required by subroutine NOISP.

(Note that the first line of the table is not used by NOISP)

	.00000	.00001	.00002	.00003	.00004	.00005	.00006	.00007	.00008	.00009
0.0	8.529781	4.261710	4.103169	4.011101	3.943466	3.889999	3.845715	3.807865	3.774779	3.745364
0.0001	3.718866	3.694745	3.672596	3.652113	3.633056	3.615232	3.598487	3.582693	3.567745	3.553555
0.0002	3.540045	3.527151	3.514819	3.502998	3.491649	3.480731	3.470213	3.460065	3.450262	3.440779
0.0003	3.431596	3.422694	3.414055	3.405663	3.397504	3.389565	3.381835	3.374301	3.366955	3.359785
0.0004	3.352784	3.345944	3.339256	3.332716	3.326314	3.320045	3.313906	3.307887	3.301987	3.296200
0.0005	3.290520	3.284945	3.279469	3.274091	3.268806	3.263610	3.258502	3.253477	3.248532	3.243666
0.0006	3.238875	3.234158	3.229510	3.224933	3.220423	3.215976	3.211592	3.207269	3.203004	3.198797
0.0007	3.194647	3.190551	3.186507	3.182515	3.178573	3.174681	3.170835	3.167035	3.163280	3.159571
0.0008	3.155904	3.152278	3.148694	3.145150	3.141644	3.138179	3.134748	3.131354	3.127997	3.124676
0.0009	3.121387	3.118132	3.114909	3.111719	3.108560	3.105432	3.102334	3.099265	3.096225	3.093214
0.0010	3.090230	3.087274	3.084344	3.081441	3.078563	3.075711	3.072884	3.070080	3.067301	3.064545
0.0011	3.061811	3.059103	3.056415	3.053749	3.051105	3.048481	3.045879	3.043298	3.040735	3.038194
0.0012	3.035671	3.033168	3.030684	3.028216	3.025769	3.023340	3.020928	3.018534	3.016156	3.013796
0.0013	3.011453	3.009125	3.006814	3.004519	3.002239	2.999975	2.997727	2.995494	2.993275	2.991071
0.0014	2.988881	2.986706	2.984544	2.982397	2.980263	2.978143	2.976035	2.973942	2.971860	2.969792
0.0015	2.967736	2.965693	2.963663	2.961644	2.959638	2.957643	2.955660	2.953688	2.951728	2.949779
0.0016	2.947842	2.945914	2.943998	2.942094	2.940199	2.938315	2.936441	2.934578	2.932725	2.930882
0.0017	2.929049	2.927225	2.925411	2.923607	2.921812	2.920027	2.918250	2.916484	2.914725	2.912976
0.0018	2.911237	2.909505	2.907783	2.906068	2.904363	2.902665	2.900977	2.899296	2.897624	2.895959
0.0019	2.894302	2.892654	2.891013	2.889380	2.887755	2.886138	2.884527	2.882925	2.881330	2.879742
0.0020	2.878160	2.876587	2.875021	2.873461	2.871908	2.870362	2.868823	2.867291	2.865766	2.864247
0.0021	2.862736	2.861230	2.859731	2.858238	2.856751	2.855271	2.853797	2.852329	2.850867	2.849412
0.0022	2.847962	2.846519	2.845081	2.843650	2.842223	2.840802	2.839389	2.837979	2.836576	2.835178
0.0023	2.833787	2.832399	2.831018	2.829642	2.828272	2.826905	2.825546	2.824191	2.822842	2.821496
0.0024	2.820157	2.818823	2.817493	2.816169	2.814850	2.813535	2.812225	2.810920	2.809620	2.808324
0.0025	2.807034	2.805747	2.804465	2.803188	2.801916	2.800649	2.799384	2.798125	2.796871	2.795621
0.0026	2.794375	2.793134	2.791897	2.790664	2.789434	2.788210	2.786990	2.785774	2.784562	2.783354
0.0027	2.782149	2.780950	2.779754	2.778562	2.777373	2.776190	2.775010	2.773833	2.772660	2.771492
0.0028	2.770327	2.769166	2.768007	2.766853	2.765703	2.764558	2.763414	2.762276	2.761140	2.760007
0.0029	2.758879	2.757753	2.756632	2.755513	2.754398	2.753287	2.752179	2.751075	2.749973	2.748876
0.0030	2.747781	2.746690	2.745602	2.744516	2.743435	2.742356	2.741282	2.740210	2.739141	2.738074
0.0031	2.737012	2.735952	2.734896	2.733842	2.732792	2.731744	2.730700	2.729658	2.728620	2.727584
0.0032	2.726551	2.725521	2.724494	2.723470	2.722448	2.721431	2.720415	2.719402	2.718392	2.717384

	.00000	.00001	.00002	.00003	.00004	.00005	.00006	.00007	.00008	.00009
0.0033	2.716380	2.715379	2.714379	2.713383	2.712389	2.711398	2.710410	2.709424	2.708442	2.707460
0.0034	2.706483	2.705507	2.704535	2.703565	2.702597	2.701632	2.700669	2.699709	2.698751	2.697797
0.0035	2.696844	2.695893	2.694946	2.694000	2.693057	2.692117	2.691178	2.690243	2.689309	2.688377
0.0036	2.687449	2.686522	2.685598	2.684676	2.683757	2.682839	2.681924	2.681011	2.680100	2.679192
0.0037	2.678286	2.677382	2.676479	2.675580	2.674683	2.673787	2.672894	2.672002	2.671114	2.670226
0.0038	2.669342	2.668459	2.667579	2.666699	2.665823	2.664948	2.664076	2.663206	2.662337	2.661470
0.0039	2.660606	2.659744	2.658883	2.658025	2.657168	2.656314	2.655460	2.654610	2.653762	2.652914
0.0040	2.652069	2.651226	2.650385	2.649546	2.648708	2.647872	2.647038	2.646207	2.645376	2.644548
0.0041	2.643722	2.642897	2.642074	2.641253	2.640433	2.639616	2.638800	2.637986	2.637173	2.636362
0.0042	2.635554	2.634747	2.633942	2.633138	2.632336	2.631536	2.630736	2.629939	2.629144	2.628350
0.0043	2.627559	2.626768	2.625979	2.625192	2.624407	2.623623	2.622841	2.622060	2.621281	2.620503
0.0044	2.619727	2.618953	2.618180	2.617409	2.616639	2.615871	2.615105	2.614340	2.613576	2.612814
0.0045	2.612054	2.611295	2.610538	2.609781	2.609027	2.608274	2.607522	2.606772	2.606024	2.605277
0.0046	2.604531	2.603786	2.603044	2.602303	2.601563	2.600824	2.600087	2.599352	2.598618	2.597884
0.0047	2.597153	2.596423	2.595694	2.594967	2.594241	2.593516	2.592793	2.592071	2.591351	2.590631
0.0048	2.589913	2.589197	2.588482	2.587768	2.587055	2.586344	2.585634	2.584926	2.584218	2.583512
0.0049	2.582808	2.582104	2.581402	2.580701	2.580001	2.579303	2.578605	2.577909	2.577214	2.576521
0.0050	2.575829	2.575138	2.574449	2.573760	2.573072	2.572387	2.571701	2.571018	2.570335	2.569654
0.0051	2.568974	2.568295	2.567617	2.566941	2.566265	2.565591	2.564919	2.564246	2.563576	2.562906
0.0052	2.562238	2.561570	2.560905	2.560239	2.559575	2.558912	2.558251	2.557590	2.556931	2.556273
0.0053	2.555615	2.554959	2.554305	2.553651	2.552998	2.552346	2.551696	2.551046	2.550398	2.549750
0.0054	2.549104	2.548459	2.547814	2.547172	2.546529	2.545888	2.545248	2.544609	2.543971	2.543334
0.0055	2.542699	2.542064	2.541430	2.540797	2.540165	2.539535	2.538905	2.538276	2.537648	2.537022
0.0056	2.536396	2.535770	2.535147	2.534524	2.533902	2.533281	2.532661	2.532043	2.531425	2.530808
0.0057	2.530192	2.529577	2.528963	2.528350	2.527738	2.527126	2.526516	2.525907	2.525298	2.524691
0.0058	2.524084	2.523479	2.522874	2.522270	2.521667	2.521066	2.520464	2.519864	2.519265	2.518666
0.0059	2.518069	2.517472	2.516877	2.516282	2.515689	2.515096	2.514503	2.513912	2.513322	2.512733
0.0060	2.512144	2.511557	2.510969	2.510384	2.509798	2.509214	2.508631	2.508048	2.507466	2.506886
0.0061	2.506306	2.505727	2.505148	2.504571	2.503994	2.503418	2.502843	2.502269	2.501696	2.501123
0.0062	2.500551	2.499981	2.499411	2.498841	2.498273	2.497705	2.497138	2.496572	2.496007	2.495442
0.0063	2.494879	2.494316	2.493754	2.493193	2.492632	2.492072	2.491513	2.490955	2.490398	2.489841
0.0064	2.489285	2.488730	2.488176	2.487622	2.487070	2.486518	2.485967	2.485416	2.484866	2.484317
0.0065	2.483768	2.483221	2.482675	2.482128	2.481584	2.481039	2.480494	2.479952	2.479409	2.478868
0.0066	2.478327	2.477787	2.477247	2.476708	2.476171	2.475634	2.475097	2.474561	2.474026	2.473491

	.00000	.00001	.00002	.00003	.00004	.00005	.00006	.00007	.00008	.00009
0.0067	2.472958	2.472425	2.471892	2.471360	2.470829	2.470299	2.469769	2.469240	2.468713	2.468184
0.0068	2.467658	2.467132	2.466606	2.466082	2.465558	2.465034	2.464512	2.463989	2.463469	2.462948
0.0069	2.462427	2.461908	2.461390	2.460871	2.460354	2.459837	2.459321	2.458805	2.458291	2.457776
0.0070	2.457263	2.456750	2.456238	2.455727	2.455215	2.454705	2.454196	2.453687	2.453178	2.452670
0.0071	2.452164	2.451657	2.451151	2.450646	2.450141	2.449637	2.449134	2.448631	2.448130	2.447628
0.0072	2.447126	2.446627	2.446127	2.445628	2.445129	2.444632	2.444135	2.443638	2.443142	2.442646
0.0073	2.442152	2.441657	2.441164	2.440671	2.440178	2.439687	2.439196	2.438704	2.438214	2.437725
0.0074	2.437236	2.436748	2.436260	2.435773	2.435287	2.434800	2.434315	2.433830	2.433346	2.432862
0.0075	2.432379	2.431896	2.431414	2.430933	2.430451	2.429972	2.429492	2.429012	2.428534	2.428056
0.0076	2.427578	2.427101	2.426624	2.426148	2.425673	2.425199	2.424725	2.424251	2.423778	2.423305
0.0077	2.422832	2.422361	2.421890	2.421420	2.420950	2.420481	2.420012	2.419543	2.419075	2.418608
0.0078	2.418141	2.417675	2.417210	2.416744	2.416280	2.415815	2.415352	2.414888	2.414426	2.413964
0.0079	2.413503	2.413041	2.412581	2.412121	2.411661	2.411202	2.410744	2.410286	2.409828	2.409371
0.0080	2.408916	2.408460	2.408004	2.407549	2.407094	2.406640	2.406187	2.405734	2.405281	2.404829
0.0081	2.404378	2.403927	2.403477	2.403027	2.402576	2.402127	2.401679	2.401231	2.400784	2.400336
0.0082	2.399890	2.399444	2.398997	2.398553	2.398108	2.397663	2.397220	2.396776	2.396334	2.395891
0.0083	2.395449	2.395008	2.394567	2.394126	2.393686	2.393247	2.392807	2.392368	2.391931	2.391493
0.0084	2.391055	2.390618	2.390182	2.389747	2.389311	2.388876	2.388441	2.388007	2.387573	2.387140
0.0085	2.386707	2.386275	2.385843	2.385411	2.384981	2.384550	2.384120	2.383691	2.383261	2.382833
0.0086	2.382404	2.381976	2.381549	2.381122	2.380695	2.380269	2.379843	2.379417	2.378993	2.378569
0.0087	2.378144	2.377721	2.377297	2.376875	2.376452	2.376031	2.375609	2.375188	2.374767	2.374347
0.0088	2.373927	2.373507	2.373089	2.372670	2.372252	2.371835	2.371417	2.371000	2.370584	2.370168
0.0089	2.369752	2.369337	2.368922	2.368507	2.368093	2.367680	2.367267	2.366854	2.366442	2.366030
0.0090	2.365618	2.365207	2.364796	2.364386	2.363976	2.363565	2.363156	2.362747	2.362339	2.361931
0.0091	2.361524	2.361116	2.360709	2.360303	2.359897	2.359491	2.359086	2.358681	2.358276	2.357872
0.0092	2.357469	2.357065	2.356662	2.356259	2.355857	2.355455	2.355054	2.354652	2.354252	2.353851
0.0093	2.353452	2.353052	2.352654	2.352254	2.351855	2.351458	2.351060	2.350662	2.350266	2.349869
0.0094	2.349473	2.349077	2.348681	2.348287	2.347892	2.347497	2.347103	2.346709	2.346316	2.345923
0.0095	2.345531	2.345139	2.344747	2.344355	2.343964	2.343574	2.343183	2.342793	2.342402	2.342013
0.0096	2.341624	2.341236	2.340848	2.340460	2.340072	2.339684	2.339298	2.338911	2.338525	2.338140
0.0097	2.337754	2.337369	2.336984	2.336599	2.336215	2.335832	2.335448	2.335065	2.334682	2.334300
0.0098	2.333918	2.333536	2.333155	2.332774	2.332393	2.332013	2.331633	2.331253	2.330873	2.330495
0.0099	2.330116	2.329738	2.329360	2.328982	2.328605	2.328228	2.327851	2.327475	2.327099	2.326723
0.0100	2.326346									

LISTING OF SUBROUTINE HAST.

```
      SUBROUTINE HAST(RV,ANS)
      T = -ALOG(RV)
      T2 = T + T
      T = SQRT(T2)
      ANS = T - (2.515517 + .802853*T + .010328*T2) /
     1     (1.0 + T*(1.432788 + .189269*T + .001308*T2))
      RETURN
      END
```


Listing of subroutine HAST as required by subroutine NORSP.

Appendix B.

This appendix lists the PL/1 implementation of CHAID that was written for the IBM S/370 Model 145 computer at the University of the Witwatersrand. The program consists of the following procedures:

1. CHANAL — A sample main program that assembles the necessary options and data required by subsequent procedures.
2. GET — A sample program for reading the input data.
3. INITIAL, ACCEPT, ANALYSE — with entry points — the procedures that perform the analysis.
4. FLOSTEP — the procedure that performs the detailed analysis of monotonic and floating predictors.
5. FRESTEP — the procedure that performs the detailed analysis of a free predictor.
6. CHI — the procedure that calculates the significance of a given chi-square variate.
7. PRINTAB — the procedure that prints the cross tabulation tables.
8. TABSIG — the procedure that calculates the significance of a contingency table.

The files used by the program are:

1. SYSPRINT — general output file
2. SUMMARY — the summary of the analysis
3. SPLIT — the detailed analysis of each predictor
4. TAB — the cross tabulation tables
5. DATUM — a temporary workfile used by the program. Its record length is 2 x (Number of predictors + 1)
6. INPUT — the input data read by procedure GET

/B 2 ...

7. SYSIN the specifications for the analysis, read by the sample main program in PL/1 LIST format.

a) Information about the dependent variable.

(i) An alphanumeric description.

(ii) Number of categories.

(iii) Alphanumeric description of first category.

(iv) " " " second

etc.

b) Number of predictors.

c) Information about each predictor.

(i) An alphanumeric description.

(ii) Its type: Monotonic ('M')
Free ('F')
Floating ('n', where n is the number of the floating category).

(iii) Whether it may be used to split the data
0 = Yes, 1 = No.

(iv) Number of categories.

(v) List of category symbols as an alphanumeric string.

(vi) Significance level for merging.

(vii) Significance level for splitting.

d) The program options.

(i) Type of cross tabulation:
0 = Row percentages,
1 = Overall percentages,
2 = Numbers per cell.

(ii) Print option.
0 = Print cross tabulation before and after each step,
1 = Only print cross tabulations before each step

(iii) Minimum number in a group for it to be analysed. 0 = Default of 50.

(iv) Minimum significance level to subdivide a group. 0 = Default of 5%.

(v) Minimum Bonferroni significance level to subdivide a group. 0 = Default of 5%.

(vi) Maximum number of groups formed for analysis to continue. 0 = Default of 10.
(Note that the last subdivision may cause this maximum to be exceeded.)

/B 4 ...

```
        CHANAL:    PROC OPTIONS (MAIN);

STMT LEVEL NEST
  1                CHANAL:    PROC OPTIONS (MAIN);
                     /*
                       THIS IS A SAMPLE MAIN PROGRAM THAT ASSEMBLES THE NECESSARY OPTIONS
                       AND DATA REQUIRED BY SUBSEQUENT PROCEDURES.                          */
                       ITS INPUT IS READ IN PL/I "LIST" FORMAT
  2     1          DCL (DEP CHAR(78) VAR, DEPNAMES (NDC) CHAR (20) VAR CTL,
                    MOER CHAR(5),  YN CHAR(3),  IV(NV) CTL,  MOPTION(6),
                    1 PRED CTL,  2 IUSE(NV),  2 MAXCAT(NV),  2 VARNAMES(NV) CHAR(50) VAR,
                      2 SIGMERG(NV),  2 SIGSPLT(NV),  2 NOFL(NV),
                      2 MF(NV) CHAR(1),   2 CATNAMES(NV), 3 CN CHAR(16) VAR;
                     /*
                       OBTAIN THE NAME OF THE DEPENDENT VARIABLE AND ITS NUMBER OF
                       CATEGORIES.                                                        */
  3     1          GET LIST(DEP,NDC);
  4     1          PUT EDIT(DEP,' HAS',NDC,' CATEGORIES CALLED:') (A,A,F(3),A);
                     /*
                       OBTAIN THE NAMES OF THE CATEGORIES OF THE DEPENDENT VARIABLE.      */
  5     1          ALLOCATE DEPNAMES;  GET LIST(DEPNAMES);  PUT EDIT(DEPNAMES) (SKIP,A);
  6     1          GET LIST(NV);
  7     1          PUT EDIT('THERE ARE',NV,' PREDICTORS AS FOLLOWS') (SKIP(3),A,F(3),A);
  8     1          ALLOCATE PRED;
  10    1          PUT EDIT('TYPE','NO. OF CATEGORIES','CATEGORY SYMBOLS','USE?','MERGE?  SPLIT?'
  11    1            ,'NAME') (SKIP(2),A,COL(9),A,COL(30),A,COL(49),A,COL(55),A,COL(72),A);
                     /*
                       OBTAIN FOR EACH PREDICTOR
                         1.  ITS NAME
                         2.  ITS TYPE - MONOTONIC, FREE OR FLOAT
                         3.  WHETHER IT MAY BE USED IN THE ANALYSIS
                         4.  HOW MANY CATEGORIES IT HAS
                         5.  SYMBOLS FOR THE CATEGORIES
                         6.  SIGNIFICANCE LEVEL FOR MERGING                             */
                         7.  SIGNIFICANCE LEVEL FOR SPLITTING
  12    1          DO I = 1 TO NV;    NOFL(I) = 0;
  14    1    1     GET LIST(VARNAMES(I),MF(I),IUSE(I),MAXCAT(I),CN(I),SIGMERG(I),SIGSPLT(I));
  15    1    1     IF MF(I) = '*' THEN MOER='MONO'; ELSE IF MF(I) = 'F' THEN MOER = 'FREE';
  19    1    1                      ELSE DO; MOER = 'FLOAT'; NOFL(I)=MF(I);  MF(I)='*'; END;
  24    1    1     IF IUSE(I) = 0 THEN YN = 'MAY';  ELSE YN = 'NOT';
                     /* CHECK NAMES.    */                                               */
                     /*
                       TRUNCATE SYMBOL LIST IF TOO LONG
  27    1    1     IF LENGTH(CN(I)) ¬= MAXCAT(I) THEN  IF LENGTH(CN(I)) > MAXCAT(I)
  29    1    1        THEN CN(I) = SUBSTR(CN(I),1,MAXCAT(I));
  30    1    1        ELSE CN(I) = CN(I) || SUBSTR((15)' ',1,MAXCAT(I)-LENGTH(CN(I)));
                                                                                         */
                     /*
                       ASSIGN HEXADECIMAL NUMBER IF TOO SHORT
  31    1    1     DO J = 1 TO MAXCAT(I);   IF SUBSTR(CN(I),J,1) = ' ' THEN
  33    1    2        SUBSTR(CN(I),J,1) = SUBSTR('123456789ABCDEFG',J,1);      END;
  35    1    1     PUT EDIT(MOER,MAXCAT(I),CN(I),YN,SIGMERG(I),SIGSPLT(I),VARNAMES(I)
                       (SKIP,A,COL(9),F(10),COL(31),A,COL(49),A,COL(55),F(6,2),COL(63),F(6,2),
                          COL(72),A);
  36    1    1       IF NOFL(I) ¬= 0 THEN PUT EDIT (NOFL(I)) (SKIP(0),COL(7),F(2));
  38    1    1     END;
```

```
        CHANAL:    PROC OPTIONS (MAIN);

STMT LEVEL NEST                                                                          */
                     /*
                       OBTAIN OPTIONS FOR THIS RUN
  39    1          GET LIST(MOPTION);
  40    1          PUT EDIT(('OPTION',I,' IS',MOPTION(I)  DO I = 1 TO DIM(MOPTION,1)))
                       (SKIP,A,F(3),A,F(5)) SKIP(2);                                     */
  41    1          CALL INITIAL (NV,NDC,MAXCAT);
                     /*  ALLOCATE SPACE FOR THE PREDICTORS                                */
  42    1          ALLOCATE IV;
                     /*
                       OBTAIN DATA USING PROCEDURE "GET"
  43    1    1     LL:  CALL GET (IDEP,IV,LAST);   IF LAST = 1 THEN GO TO CALC;
  46    1          CALL ACCEPT (IDEP,IV);  GO TO LL;                                    */
                     /*
                       PERFORM ANALYSIS
  48    1          CALC: CALL ANALYSE(MOPTION,DEPNAMES,VARNAMES,CATNAMES,NF,IUSE,SIGMERG,
                                        SIGSPLT,NOFL);                                   */
                     /*
                       ANALYSIS COMPLETED
  49    1          PUT EDIT ('THIS IS THE VERY END OF THE ANALYSIS. G V KASS 1974')(SKIP(3),A);
  50    1          STOP;                    END CHANAL;
  51    1
```

```
        GET:  PROC (IDEP,IV,LAST);

STMT LEVEL NEST
  1              GET:  PROC (IDEP,IV,LAST);
                   /*
                     A SAMPLE PROGRAM FOR READING THE DATA
                     THIS ROUTINE ASSUMES THE DATA IS IN TWO COLUMN FIELDS, THE FIRST
                     ITEM BEING THE DEPENDENT VARIABLE                              */
  2     1          DCL  IV(*);
  3     1          ON ENDFILE (INPUT) GO TO FN;
  5     1          GET FILE (INPUT) EDIT (IDEP,IV) (F(2));
  6     1          LAST = 0;  RETURN;
                   /*
                     AT THE END OF THE DATA SET THE FOLLOWING CODING IS USED        */
  8     1        FN:  LAST = 1;  RETURN;  END GET;
```

```
        INITIAL:  PROC (NOVAR,NCAT,MAXCAT);

STMT LEVEL NEST
  1              INITIAL:  PROC (NOVAR,NCAT,MAXCAT);
                   /*
                     THIS ROUTINE PERFORMS THE ANALYSIS OF THE DATA, AND CONTAINS A
                     NUMBER OF ENTRY POINTS                                         */
                   /*
                     THIS SECTION PERFORMS THE NECESSARY ALLOCATION AND INITIALIZATION
                     OF DATA STORAGE IN THE COMPUTER                                */
  2     1          DCL  MAXCAT(*),  MC(NV) CTL,
                     1 TABLE(NV) CTL,  2 MAT(MNT,NCT) INIT ((NV*MNT*NCT)0),
                     (MN,MNT,NC,NCT,NV) STATIC;
  3     1          DCL  1 DATA CTL,  2 ID,  2 IV(NV);
  4     1          NV=NOVAR;  NC=NCAT;  NCT=NC+1;  ALLOCATE MC;    MN=1;
  9     1          ALLOCATE DATA;
 10     1          DO I = 1 TO NOVAR;  K = MAXCAT(I);
 12     1    1       IF K <= 0 THEN K = 1;    IF K > 16 THEN K = 16;
 16     1    1       MC(I) = K;  IF K > MN THEN MN = K;      END;
 20     1          MNT = MN+1;   ALLOCATE TABLE;    RETURN;
 23     1        ACCEPT:  ENTRY (IDEP,IVAR);
                   /*
                     THIS SECTION ACCEPTS THE INPUT DATA CASE BY CASE, AND CALCULATES THE
                     CONTINGENCY TABLES AND STORES THE DATA FOR FUTURE REFERENCE     */
 24     1          DCL IVAR(*);
 25     1          ID = IDEP;  IF ID<=0 THEN ID=1;   IF ID>NC THEN ID=NC;
 30     1          DO I = 1 TO NV;  K = IVAR(I);  IF K<=0 THEN K=1;  IF K>MC(I) THEN K=MC(I);
 36     1    1       IV(I) = K;  MAT(I,K,ID) = MAT(I,K,ID) + 1;     END;
 39     1          WRITE FILE (DATUM) FROM (DATA);   RETURN;
 41     1        ANALYSE:  ENTRY(NOPTION,DEPNAMES,VARNAMES,CATNAMES,NF,NUSE,SIGMERG,
                            SIGSPLT,NOFL);
                   /*
                     THIS SECTION PERFORMS THE DATA ANALYSIS
                   /*
                     FIRSTLY THE OPTIONS ARE INTERPRETED AND DEFAULTS SET IF NECESSARY     */
                   /*   NOPTION(1)  =0 PRINT TABLES AS ROW %,  =1 AS #,  ELSE AS NUMBERS.
                   /*   NOPTION(2)  =0 PRINT TABLES AT EACH STEP,  OTHERWISE NOT.
                   /*   NOPTION(3)  MINIMUM NUMBER IN A GROUP.  <=20 TAKE DEFAULT OF 50.
                   /*   NOPTION(4)  MINIMUM % SIGNIFICANCE TO SPLIT A GROUP.  =0 TAKE AS 5%
                   /*   NOPTION(5)  MINIMUM % BONF SIGNIF. TO SPLIT A GROUP.  =0 TAKE AS 5%
                   /*   NOPTION(6)  MAXIMUM NUMBER OF GROUPS ALL TOLD.  =0 TAKE AS 10.
                                                                                    */
 42     1          DCL  SUMMARY PRINT,  SPLIT PRINT;
 43     1          DCL  NOPTION(*), DEPNAMES(*) CHAR(*)VAR,  VARNAMES(*) CHAR(*)VAR,
                       SIGMERG(*),  SIGSPLT(*),  NOFL(*),
                     NF(*) CHAR(1),  NUSE(*),  TITLE CHAR(500) VAR EXT,  PERC CHAR(1),
                     1 CATNAMES(*),  2 CN CHAR(4) VAR,  IPOINT(NOPTION(5)*16) CTL,
                    1 RES, 2 SIGMID, 2 BONFD, 2 NOGD, 2 NOSD(32) INIT (1,1,1,2,1,3,1,4,1,5,
                       1,6,1,7,1,8,1,9,1,10,1,11,1,12,1,13,1,14,1,15,1,16),
                     1 RESULT, 2 SIGMIN, 2 BONF, 2 NOG, 2 NO(32),     TAB PRINT,  NOSDTMP(32),
                     ICA(300),  NOGROUPS INIT(1),  MAXPT INIT(2),  ICT(NCT) CTL;
 44     1          MAXICA = 300;
                   /*
                     HEADINGS ARE PREPARED FOR THE VARIOUS OUTPUT FILES              */
 45     1          ON ENDPAGE (SUMMARY)  PUT FILE (SUMMARY) EDIT
                     ('SUMMARY OF RESULTS',TITLE) (A,SKIP,A)
                     (' NO.',' NAME',' SIGNIF (%)',' BONF=SIG (%)',' NUMBER OF GROUPS.')
                     (SKIP(2),A,COL(11),A,COL(55),A,COL(67),A,COL(80),A) PAGE;
 47     1          TITLE='ANALYSIS OF TOTAL GROUP.';   NOSDTMP = NOSD;
```

```
INITIAL:  PROC (NOVAR,NCAT,MAXCAT);

STMT LEVEL NEST

                    /*                                                    */
                      OPTIONS ARE ANALYSED
 49     1           PUT EDIT ('OPTIONS IN EFFECT:','TABLES PRINTED ') (SKIP(3),A,SKIP,A);
 50     1           IF NOPTION(1)=0 THEN DO; PERC='P'; PUT EDIT('WITH ROW PERCENTAGES')(A); END;
 58     1             ELSE IF NOPTION(1) = 1
 54     1               THEN DO; PERC='P'; PUT EDIT('AS OVERALL PERCENTAGES')(A); END;
 60     1               ELSE DO; PERC='N'; PUT EDIT('GIVING TOTLA NUMBER WITHIN EACH CELL')
 62     1     1                                        (A);   END;
 64     1           IF NOPTION(2)=0
 65     1             THEN PUT EDIT('TABLES WILL BE PRINTED BEFORE & AFTER EVERY STEP')(SKIP,A);
 66     1             ELSE PUT EDIT('TABLES WILL BE PRINTED ONLY BEFORE EACH STEP')(SKIP,A);
 67     1           IF NOPTION(3) < 20 THEN NOPTION(3) = 50;
 69     1           PUT EDIT('A GROUP WITH LESS THAN',NOPTION(3),' CASES WILL NOT BE ANALYSED')
                          (SKIP,A,F(4),A);
 70     1           IF NOPTION(4) <= 0 THEN NOPTION(4) = 5;
 72     1           PUT EDIT('A GROUP WILL BE SPLIT IF IT IS SIGNIFICANT AT THE',NOPTION(4),
                      '% LEVEL') (SKIP,A,F(3),A);
 73     1           IF NOPTION(5) <= 0 THEN NOPTION(5) = 5;
 75     1           PUT EDIT('A GROUP WILL BE SPLIT IF IT IS BONFERRONI SIGNIFICANT AT THE',
                      NOPTION(5),'% LEVEL') (SKIP,A,F(3),A);
 76     1           IF NOPTION(6) <= 0 THEN NOPTION(6) = 10;
 78     1           PUT EDIT('THE PROGRAM WILL TERMINATE WHEN',NOPTION(6),
                      ' OR MORE GROUPS HAVE BEEN FORMED OR ANALYSED')(SKIP,A,F(3),A);
                                                                          */
                    /*
                      INITIALIZE POINTERS
 79     1           ALLOCATE IPOINT, ICT;  IPOINT(1)=1;  ICA(1)=0;  IFROM=1;
                    /*
                      PREPARE FOR ANALYSIS BY COMPLETING THE CONTINGENCY TABLES THAT HAVE   */
                      BEEN CREATED
 83     1           ANALYSE: CLOSE FILE (DATUM);
                      /* FILL IN COLUMN TOTALS.  */
 84     1           DO I=1 TO NCT; IS=0; DO J=1 TO MN; IS=IS+MAT(I,J,I); END; ICT(I)=IS; END;
 91     1           IGRTOT = SUM(ICT);
                      /* ROW TOTALS.  */
 92     1           GL: DO I=1 TO NV;  MAT(I,MNT,*) = ICT;  MAT(I,MNT,NCT) = IGRTOT;
 95     1     1       DO J=1 TO MN; IS=0;  DO K=1 TO NC; IS=IS+MAT(I,J,K);  END;
100     1     2         MAT(I,J,NCT)=IS;  END;                  END GL;
                                                                          */
                    /*
                      ORIGINAL CONTINGENCY TABLES ARE PRINTED
103     1           TITLE = TITLE || ' << BEFORE GROUPING >>';     NOSD = NOSDTMP;
105     1           PUT FILE(TAB) EDIT ('G V KASS. 1974. V.A.',TITLE) (A,SKIP,A) PAGE;
106     1           DO I=1 TO NV;   NOGD=MC(I);
108     1     1       CALL TABSIG(TABLE(I),X,DF,SIGNI);
109     1     1       CALL PRINTAB(TABLE(I),CN(I),VARNAMES(I),RES,DEPNAMES,PERC,SIGNI,TAB);
110     1     1                 END;
111     1           TITLE = SUBSTR(TITLE,1,LENGTH(TITLE)-23);
                                                                          */
                    /*
                      PRINT HEADINGS IN VARIOUS PRINT-FILES
112     1           IF NOPTION(2) = 0 THEN
113     1           PUT FILE (TAB) EDIT ('G V KASS. 1974. V.A.',TITLE) (A,SKIP,A) PAGE;
114     1           SIGNAL ENDPAGE (SUMMARY);
115     1           PUT FILE (SPLIT) EDIT('G V KASS. 1974. V.A.',TITLE)(A,SKIP,A)PAGE;
```

```
INITIAL:  PROC (NOVAR,NCAT,MAXCAT);

STMT LEVEL NEST

                    /*
                      THIS LONG LOOP THAT FOLLOWS ANALYSES A SUBGROUP AND PERFORMS THE
                      NECESSARY DIVISION OF THE DATA ACCORDING TO THE RESULT OF THE          */
                      ANALYSIS, ON EACH ITERATION
116     1           LOOP:  SIGMIND = 100;  BONFD = 1;                     */
                    /*
                      ANALYSE A SUBGROUP
118     1           DO I = 1 TO NV;
119     1     1       IF ME(I) = 'M'                                      */
                      /*
                        "FLOSTEP" IS USED BOTH FOR FLOATING AND MONOTONIC PREDICTORS
120     1     1       THEN CALL FLOSTEP(TABLE(I),CN(I),VARNAMES(I),RESULT,SIGMERG(I),
                            SIGSPLT(I),NOFL(I));
121     1     1       ELSE CALL FRESTEP(TABLE(I),CN(I),VARNAMES(I),RESULT,SIGMERG(I),SIGSPLT(I));
                                                                          */
                      /*
                        SUMMARIZE THE RESULTS
122     1     1       CALL SUMMAR;                                        */
                      /*
                        FIND THE BEST WAY TO SPLIT THE DATA
                      IF IUSE(I) = 0 THEN IF SIGMIN*BONF < SIGMIND*BONFD THEN
123     1     1         DO; RES = RESULT;  IB = I;  END;
125     1     1                                                           */
129     1     1       IF NOPTION(2) = 0 THEN
                      /*
                        TABULATE RESULTS
130     1     1       CALL PRINTAB(TABLE(I),CN(I),VARNAMES(I),RESULT,DEPNAMES,PERC,SIGMIN,TAB);
131     1     1     END;                                                  */
132     1           IF SIGMIND = 100 THEN DO;
                      /*
                        NO SPLIT CAN BE USED
134     1     1     PUT FILE(SPLIT) EDIT('CANNOT EXPLAIN ANYTHING.')(SKIP(3),A);
135     1     1     PUT FILE(SUMMARY) EDIT('NO PREDICTION IS POSSIBLE.')(SKIP(2),A);
136     1     1                       GO TO NEXTGROUP;    END;
                                                                          */
                      /*
                        THE BEST PREDICTOR IS FOUND
138     1           PUT FILE(SPLIT) EDIT('BEST TO USE VAR',IB,VARNAMES(IB),' TO GIVE',NOGD,
                      ' NEW GROUPS') (SKIP(3),A,F(3),X(1),A,A,F(2),A);
139     1           PUT FILE(SUMMARY) EDIT ('CHARACTERISTICS OF THE BEST PREDICTOR')(SKIP(2),A);
140     1           I = IB;  RESULT = RES;  CALL SUMMAR;
                                                                          */
                      /*
                        TEST IF BEST PREDICTOR IS SIGNIFICANT ENOUGH TO SUBDIVIDE THE DATA
                      IF (SIGMIND > NOPTION(4)) | (SIGMIND*BONFD > NOPTION(5))  THEN DO;
143     1             PUT FILE (SPLIT) EDIT ('THIS DIVISION IS NOT SIGNIFICANT ENOUGH.')
145     1     1             (SKIP(2),A);
                      PUT FILE (SUMMARY) EDIT('THIS PREDICTOR IS NOT SIGNIFICANT.')(SKIP(2),A);
146     1     1               GO TO NEXTGROUP;   END;
147     1     1                                                           */
                      /*
                        DATA WILL BE SUBDIVIDED ACCORDING TO THE FOLLOWING PLAN
149     1           CALL PRINTAB(TABLE(I),CN(I),VARNAMES(I),RESULT,DEPNAMES,PERC,SIGMIN,SUMMARY);
                      /* CREATE NEW GROUP'S POINTERS.   */
150     1           IP = I;
```

```
          INITIAL:  PROC (NOVAR,NCAT,MAXCAT);

STMT LEVEL NEST

                      /*                                                          */
                        UPDATE POINTERS
151     1           DO I = 1 TO NOGO;  NOGROUPS = NOGROUPS + 1;  ITO = NOGROUPS;
154     1     1       IPL, IPF = IPOINT(IFROM);  IPOINT(ITO) = MAXPT;
156     1     1     LP1: IF ICA(IPL)=0 THEN GO TO OUL;
158     1     1       ICA(MAXPT)=ICA(IPL); IPL=IPL+1;  MAXPT=MAXPT+1;  GO TO LP1;
                      /*                                                          */
                        CHECK IF ENOUGH STORAGE HAS BEEN ALLOCATED
162     1     1     IF  MAXPT+NOSO(IP)+2 > MAXICA THEN DO;
164     1     2       PUT EDIT('1','PROGRAM STORAGE CAPACITY EXCEEDED - INCREASE <ICA>',
165     1     2           ('130'P'A')  (SKIP(3),A,SKIP(2),A,SKIP(2),A);  GO TO FINISH;  END;
                      /*                                                          */
                        DEFINE THE NEW GROUPS
                      /* ADD ONTO ICA THIS SPLIT  -  OUT OF THE LOOP   */
167     1     1     OUL:  ICA(MAXPT)=10;  ICA(MAXPT+1)=NOSO(IP);  NOIN=0;
170     1     1       DO J = 1 TO NOSO(IP);  ICA(MAXPT+J+1) = NOSO(IP+J);
172     1     2           NOIN = NOIN + MAT(IB,NOSO(IP+J),NCT);    END;
174     1     1       ICA(MAXPT+NOSO(IP)+2) = 0;  MAXPT = MAXPT + NOSO(IP) + 3;
176     1     1       IF NOINCNOPTION(3) THEN IPOINT(ITO)=-IPOINT(ITO);
178     1     1       IP = IP + NOSO(IP) + 1;    END;
                      /*                                                          */
                        FIND AND PREPARE FOR NEXT GROUP TO BE ANALYSED
180     1           NEXTGROUP:  IPOINT(IFROM) = -IPOINT(IFROM);
181     1           IF NOGROUPS >= NOPTION(6) THEN GO TO FINISH;
183     1           DO I = 1 TO NOGROUPS;  IF IPOINT(I) > 0 THEN GO TO SETUP;  END;
187     1           GO TO FINISH;
                      /*                                                          */
                        SUBGROUP TO BE ANALYSED HAS BEEN FOUND
188     1           SETUP:  IFROM = I;  IPL, IPF =IPOINT(IFROM);  TABLE = 0;
191     1           TITLE = 'ANALYSIS OF SUBGROUP ';
192     1           LP2:  ICAIPL = ICA(IPL);  IF ICAIPL = 0 THEN GO TO SETUP2;
                      /*                                                          */
                        PREPARE PAGE HEADING
195     1           TITLE=TITLE||VARNAMES(ICA(IPL))||' ('||SUBSTR(CN(ICA(IPL)),ICA(IPL+2),1);
196     1           DO J=1 TO ICA(IPL+1)-1;
197     1     1       TITLE=TITLE||','||SUBSTR(CN(ICA(IPL)),ICA(IPL+J+2),1);    END;
199     1           TITLE=TITLE||')  ';   IPL = IPL + ICA(IPL+1) + 2;  GO TO LP2;
202     1           SETUP2:  ON ENDFILE (DATUM) GO TO ENDSETUP;
                      /*                                                          */
                        PREAD DATA CHOOSING THOSE IN THIS SUBGROUP
204     1           LP3:  READ FILE (DATUM) INTO (DATA);
                      /* CHECK IF DATA IS REQUIRED,  */
205     1           IPL = IPF;
206     1           LP4:  ICAIPL = ICA(IPL);   IF ICAIPL = 0 THEN GO TO TAKE;
209     1           DO J=1 TO ICA(IPL+1); IF IV(ICAIPL)=ICA(IPL+J+1) THEN GO TO LP5;  END;
213     1           GO TO LP3;
214     1           LP5:  IPL = IPL + ICA(IPL+1) + 2;  GO TO LP4;
```

```
          INITIAL:  PROC (NOVAR,NCAT,MAXCAT);

STMT LEVEL NEST

                      /*                                                          */
                        FORM CONTINGENCY TABLE FROM THIS ITEM
216     1           TAKE:  DO I=1 TO NV;  MAT(I,IV(I),ID) = MAT(I,IV(I),ID) + 1;   END;
219     1           GO TO LP3;
                      /*                                                          */
                        ANALYSE THIS DATA
220     1           ENDSETUP:     GO TO ANALYSEP;
                      /*                                                          */
                        PRINT A SINGLE LINE SUMMARY OF THE RESULT OF ANALYSING WITH A
                        SINGLE PREDICTOR
221     1           SUMMARY:  PROC;
222     2             PUT FILE(SUMMARY) EDIT(I,VARNAMES(I)||(25)' ',SIGMIN,SIGMIN*BONF,NOG,' ')
                        (SKIP,F(3),X(1),A(47),F(13,6),F(14,8),F(3),A);
223     2             IP = 1;  DO J = 1 TO NOG;  PUT FILE(SUMMARY) EDIT('  ') (A);
226     2     1         DO K = 1 TO NO(IP);
227     2     2           PUT FILE(SUMMARY)EDIT(SUBSTR(CN(I),NO(IP+K),1))(A);  END;
229     2     1           IP = IP + NO(IP) + 1;   END;
231     2                                   END SUMMARY;
                      /*                                                          */
                        CONCLUDING CRITERION HAS BEEN REACHED -- END OF ANALYSIS
232     1           FINISH:  PUT EDIT('END OF ANALYSIS','G V KASS. 1974')(SKIP(3),A,X(50),A);
233     1             RETURN;       END INITIAL;
```

```
FLOSTEP:  PROC  (MATRS,HEADR,NAME,RESULT,SIGMERG,SIGSPLT,NOFL);

STMT LEVEL NEST
 1           FLOSTEP:  PROC  (MATRS,HEADR,NAME,RESULT,SIGMERG,SIGSPLT,NOFL);
               /*
                 THIS PROCEDURE ANALYSES IN DETAIL A FLOATING OR MONOTONIC
                 PREDICTOR ON THE BASIS OF A CONTINGENCY TABLE PROVIDED BY THE
                 CALLING PROGRAM                                              */
 2     1       DCL  1 MATRS, 2 MATR(*,*),  HEADR CHAR(*) VAR,  NAME CHAR(*) VAR,
               TITLE CHAR(500) VAR EXT,
               SPLIT PRINT,
                 SPLIT ST ENTRY (BIN FIXED, BIN FIXED),
               1 RESULT, 2 SIGMIN, 2 BONF, 2 NOG, 2 NOSR(32),
               N(17) INIT (1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17),
               1 CONTROL CTL,  2 MAT(IR,IC),  2 (E,O,CO)(IR-1,IC-1),  2 MATT(IC),
                 2 (FEA,FOA,REB,ROB,RET,ROT,OT,CCT)(IC-1),  2 (RT,IREM)(IR),
                 2 (HEAD,HEADX) CHAR(IR),  2 NOFCH CHAR(1);
                 /*
                   PROVIDE HEADINGS FOR OUTPUT                                */
 3     1       ON ENDPAGE(SPLIT) PUT FILE(SPLIT) EDIT('G V KASS. 1974. V.A.',TITLE,
                 'CONTINUED',***) (A,SKIP,A,SKIP,COL(100),A,SKIP,A) PAGE;
 5     1       IF LINENO (SPLIT) > 52 THEN SIGNAL ENDPAGE (SPLIT);
 7     1       IF NOFL = 0 THEN
 8     1       PUT FILE (SPLIT) EDIT((130)'*','MONOTONIC PREDICTOR:  ',NAME)
                 (SKIP,A,SKIP,A,A);
 9     1       ELSE
 9     1       PUT FILE (SPLIT) EDIT ((130)'*','SINGLE FLOATING PREDICTOR:  ',NAME)
                 (SKIP,A,SKIP,A,A);
10     1       BONF = 1;
11     1       NOF = 0;  NOFGR = 0;
                 /*
                   REMOVE ROWS AND COLUMNS WITH ZERO MARGINAL TOTALS           */
13     1       K = DIM(MATR,1);  L = DIM(MATR,2);  IR, IC = 1;
16     1       DO I = 1 TO K-1;  IF MATR(I,L) > 0 THEN IR = IR + 1;  END;
20     1       DO I = 1 TO L-1;  IF MATR(K,I) > 0 THEN IC = IC + 1;  END;
24     1       ALLOCATE CONTROL;  IS = 0;
26     1       DO I = 1 TO K;  IF MATR(I,L) <= 0 THEN GO TO ENDI;
29     1  1      IS=IS+1;  IREM(IS)=I;  IT=0;  SUBSTR(HEAD,IS,1)=SUBSTR(HEADR,I,1);
33     1  1    IF I = NOFL THEN DO; NOF = IS;  NOFCH = SUBSTR(HEADR,I,1); NOFGR=1; END;
38     1  1      ELSE SUBSTR(HEADX,IS-NOFGR,1) = SUBSTR(HEADR,I,1);
40     1  1      DO J = 1 TO L;  IF MATR(K,J) <= 0 THEN GO TO ENDJ;
43     1  2        IT = IT + 1;  MAT(IS,IT) = MATR(I,J);
45     1  2      ENDJ:  END;     ENDI:  END;
47     1       K = IR;  L = IC;   M = K-1;
                 /*
                   SUSPEND ANALYSIS IF LESS THAN TWO ROWS OR COLUMNS           */
50     1       IF L<=2 THEN DO; PUT FILE (SPLIT) EDIT('DEPENDENT VARIABLE IS A CONSTANT.')
53     1  1      (SKIP(2),A);   GO TO FINISHA;  END;
55     1       IF K<=2 THEN DO; PUT FILE (SPLIT) EDIT (' IS A CONSTANT.')(A);
58     1  1      GO TO FINISHA;   END;
60     1       IF NOF <= 1 THEN GO TO PREFIX;
                 /*
                   RE-ORDER SO THAT FLOATING CATEGORY IS FIRST ON LIST          */
62     1       SUBSTR(HEAD,1,1) = NOFCH;  SUBSTR(HEAD,2,IR-1) = HEADX;  MATT=MAT(NOF,*);
65     1       DO I = NOF TO 2 BY -1;  MAT(I,*) = MAT(I-1,*);  IREM(I) = IREM(I-1);  END;
69     1       MAT(1,*) = MATT;  NOF = 1;  IREM(1) = NOFL;
```

```
FLOSTEP:  PROC  (MATRS,HEADR,NAME,RESULT,SIGMERG,SIGSPLT,NOFL);

STMT LEVEL NEST
                 /*
                   CALCULATE EXPECTED VALUES                                  */
72     1       PREFIX:  X = 0;  DO I = 1 TO M;  R = 0;  DO J = 1 TO L-1;
76     1  2      E(I,J) = FLOAT(MAT(I,L)*MAT(K,J)) / MAT(K,L);  O(I,J)=MAT(I,J);
78     1  2      IF O(I,J) <= 1.0E-30 THEN O(I,J) = 1.0E-30;
80     1  2      R = R + (O(I,J)-E(I,J))**2 / E(I,J);   END;      X = X + R;    END;
84     1       DF = (M-1)*(L-2);  SIG = CHI(X,DF);
86     1        PUT FILE(SPLIT) EDIT('TABLE HAS CHI-SQUARE',X,', WITH D.F.',DF,
                   ', AND SIGNIF',SIG,'*') (SKIP,A,F(9,3),A,F(4),A,F(13,8),A);
87     1       DF = L-2;  NLOOP = 0;  MAXLOOP = 3*M;  CO = 0;  RT = MAT(*,IC);
92     1       NOFGR = 0;
                 /*
                   THIS LOOP MERGES AND SPLITS THE CATEGORIES                  */
93     1       LOOP:  IF M <= 1 THEN GO TO TRYSPLT;
95     1        IF LINENO (SPLIT) > 54 THEN SIGNAL ENDPAGE (SPLIT);
                   /*
                     TREAT AS MONOTONIC UNDER THE FOLLOWING CONDITIONS         */
97     1        IF NOF=0 | M=2 | NOFGR¬=0 THEN DO; NOFT = 0;  ISTA = 1;
                   /*
                     SUBHEADING                                                */
101    1  1        PUT FILE (SPLIT) EDIT (M,' GROUPS:',(SUBSTR(HEAD,I,1)DO I=1 TO K-1))
102    1  1            (SKIP(2),F(3),A,COL(28),16 A(6));    END;
103    1         ELSE  DO;  NOFT = NOF;  ISTA = 2;
106    1  1        PUT FILE (SPLIT) EDIT (M,' GROUPS:',(SUBSTR(HEADX,I,1)DO I=1 TO K-2))
107    1  1            (SKIP(2),F(3),A,COL(28),16 A(6));    END;
                   /* PUT IN APPROPRIATE BRACKETS.   */
108    1        ISP = 0;   DO I = 1 TO M;
110    1  1        IF NO(I) = NOFT THEN ISP = 1;
112    1  1          ELSE IF NO(I+1) ¬= NO(I)+1 THEN PUT FILE (SPLIT) EDIT('(',')')
                     (SKIP(0),COL(6*(NO(I)-ISP)+20),A,COL(6*(NO(I+1)-ISP)+18),A);
114    1  1        END;
                   /*
                   /* CALCULATE STATS FOR GROUPING  I  AND  I+ISP    */
115    1        IM = 0;  XT = 1.E60;  ISP = 1;
118    1        PUT FILE (SPLIT) EDIT ('TEST STATS FOR GROUPING:') (SKIP,A(27));
119    1        DO I = ISTA TO M-1;
120    1  1        ROT = (O(NO(I),*)  + O(NO(I+ISP),*)) / (RT(NO(I)) + RT(NO(I+ISP)));
121    1  1        REA = ROT*RT(NO(I));   REB = ROT*RT(NO(I+ISP));
123    1  1        R = SUM((O(NO(I),*)-REA)**2/REA + (O(NO(I+ISP),*)-REB)**2/REB);
124    1  1        PUT FILE (SPLIT) EDIT (R) (X(6*(NO(I+ISP)-NO(I)-ISP)),F(6,1));
                   /*
                     FIND MINIMUM MERGE STATISTIC                              */
125    1  1        IF R < XT THEN DO; XT = R;  IM = I;  IM2 = I + ISP;  END;
131    1  1      ELI:  END;
132    1         IF NOFT = 0 THEN GO TO EMERGE;
                   /*
                     COMPARE FLOATING CATEGORY WITH EACH OF THE OTHERS         */
134    1        PUT FILE (SPLIT) EDIT ('AND FOR GROUPING ',NOFCH,' WITH:')(SKIP,A,A,A);
135    1        DO I = 2 TO M;   IMT = 1;
137    1  1        ROT = (O(NO(I),*) + O(NOF,*)) / (RT(NO(I)) + RT(NOF));
138    1  1        REA = ROT*RT(NO(I));   REB = ROT*RT(NOF);
140    1  1        R = SUM((O(NO(I),*)-REA)**2/REA + (O(NOF,*)-REB)**2/REB);
141    1  1        PUT FILE (SPLIT) EDIT (R) (X(6*(NO(I+1)-NO(I)-1)),F(6,1));
142    1  1        IF R < XT THEN DO; XT = R;  IM = -I;  END;
147    1  1      ELI2:  END;
```

```
FLOSTEP:  PROC  (MATSS,HEADR,NAME,RESULT,SIGMERG,SIGSPLT,NOFL);

STMT LEVEL NEST

                                                                          */
                  /*
                    ANALYSE THE BEST MERGE
148      1        EMERGE:   SIGT = CHI(XT,DF);
149      1         PUT FILE (SPLIT) EDIT('MIN STAT IS',XT,',  TO MERGE (',(SKIP(2),A,F(8,4),A);
150      1         IF IM > 0 THEN DO;  IG1 = NO(IM);  IG2 = NO(IM2);
154      1    1          PUT FILE (SPLIT) EDIT (SUBSTR(HEAD,IG1,IG2-IG1),',') AND (',
155      1    1               SUBSTR(HEAD,IG2,NO(IM2+1)-IG2),')')(4 A);   END;
156      1              ELSE DO;  IG1 = NO(-IM);  IG2 = NOF;  IM2 = IMT;
160      1    1               PUT FILE(SPLIT) EDIT(NOFCH,') AND (',
161      1    1                    SUBSTR(HEAD,IG1,NO(1-IM)-IG1),')')(4 A);   END;
162      1         PUT FILE (SPLIT) EDIT ('D.F.=',DF,',  SIGNIFICANCE',SIGT,'%')
                    (COL(64),A,F(4),A,F(9,4),A);
                                                                          */
                  /*
                    IF CANNOT MERGE, TRY SPLITTING
163      1         IF SIGT < SIGMERG THEN GO TO TRYSPLT;
                    /* RELOCATE POINTS IF NECESSARY   */
165      1         IF NOFGR = 0 THEN DO;  IF IM < 0 THEN NOFGR = IG1;
169      1    1                                ELSE IF IG1=NOF THEN NOFGR = IG2;
171      1    1                                     ELSE IF IG2=NOF THEN
172      1    1                                          NOFGR=IG1;
173      1    1                END;
174      1              ELSE IF IG2 = NOFGR THEN NOFGR = IG1;
                    /*
                    /*  COMBINE GROUPS IN POSITIONS  IG1  AND  IG2    */
176      1         RT(IG1) = RT(IG1) + RT(IG2);
177      1         O(IG1,*) = O(IG1,*) + O(IG2,*);
178      1         DO I = IM2 TO M;  NO(I) = NO(I+1);  END;      M = M - 1;
                    /*
                    /* REORDER IF NECESSARY    */
192      1         IF IM >= 0 THEN GO TO NLPTST;
194      1         IF IM2 = 1-IM THEN GO TO NLPTST;
196      1         SWAPEND:   SUBSTR(HEAD,IG2,IG1-IG2) = SUBSTR(HEADX,IG2);
197      1         NOFGR = NOFGR - 1;
198      1         NOF = IG1-1;    SUBSTR(HEAD,NOF,1) = NOFCH;
                                                                          */
                    /*
                      CALCULATE NEW SUBTABLES
200      1         MAT = MAT(IG2,*);  COT = CO(IG2,*);
192      1         DO I = IG2 TO IG1-2;
193      1    1      MAT(I,*) = MAT(I+1,*);  O(I,*) = O(I+1,*);  CO(I,*) = CO(I+1,*);
196      1    1      RT(I) = RT(I+1);  IREM(I) = IREM(I+1);      END;
199      1         MAT(NOF,*) = MAT;  CO(NOF,*) = COT;  IREM(NOF) = NOFL;
202      1         O(NOF,*) = O(IG1,*);  RT(NOF) = RT(IG1);
204      1         DO I = IM2 TO -IM-1;  NO(I) = NO(I) - 1;  END;
                                                                          */
                    /*
                      TEST IF PROCEDURE IS IN A LOOP -- EMERGENCY EXIT
207      1         NLPTST:  NLOOP = NLOOP + 1;  IF NLOOP > MAXLOOP THEN DO;  PUT EDIT
                    ('FORCE TERMINATION - SEEM TO BE IN A LOOP')(SKIP(2),A);
211      1    1      PUT FILE(SPLIT) EDIT('FORCE TERMINATION - IN A LOOP')(SKIP(2),A);
212      1    1      GO TO FINISH;  END;
214      1         GO TO LOOP;
```

```
FLOSTEP:  PROC  (MATFS,HEADR,NAME,RESULT,SIGMERG,SIGSPLT,NOFL);

STMT LEVEL NEST

                  /*                                                      */
                    TRY SPLITTING A COMBINED GROUP.     NO SPLITTING ATTEMPTED IF
                    SIGNIFICANCE OF MERGING LESS THAN THAT OF SPLITTING
                    SIGNIFICANCE OF MERGING LESS THAN THAT OF SPLITTING
215      1         TRYSPLT:  IF SIGMERG <= SIGSPLT THEN GO TO FINISH;
217      1         PUT FILE (SPLIT) EDIT ('M,' GROUPS:')(SKIP(2),F(3),A(2));
218      1         ISP = 1;  IMT = 0;                                    */
                    /*
                      TRY EACH GROUP
220      1         DO I = 1 TO M;
221      1    1      IF NO(I+1) = NO(I) + 1 THEN DO;  PUT FILE(SPLIT) EDIT
                      (SUBSTR(HEAD,NO(I),1))(X(ISP+2),A);  ISP = 3;  GO TO EHEAD;  END;
224      1    2      IF NO(I) = NOFGR & NO(I+1) ^= NOFGR+2 THEN GO TO SPECIAL;
227      1    1      PUT FILE (SPLIT) EDIT ('( ',(SUBSTR(HEAD,J,1)DO J = NO(I) TO NO(I+1)-1),
229      1    1       ' )') (X(ISP),A,A,(NO(I+1)-NO(I)-1) (X(5),A),A);
                                                                          */
230      1    1      ISP = 1;  GO TO EHEAD;
                      /*
                        SPECIAL CODING IF COMBINED GROUP CONTAINS THE FLOATING CATEGORY
232      1    1      SPECIAL:   IMT = I;
233      1    1       PUT FILE (SPLIT) EDIT('( ',NOFCH,(SUBSTR(HEADX,J,1)DO J=NO(I) TO NO(I+1)-2),
234      1    1        NOFCH,' )') (X(ISP),A,(NO(I+1)-NO(I)) A(6),A,A);  ISP = 1;
235      1    1      EHEAD:   END;
236      1         IM = 0;  XT = -1;  NO(M+1) = IK;  ISP = 22;
240      1         PUT FILE(SPLIT) EDIT('TEST STATS FOR SPLITTING:') (SKIP,A);
                                                                          */
                    /*
                      ANALYSE THIS COMBINED GROUP
241      1         TSLOOP:  DO I = 1 TO M;  IF NO(I+1) = NO(I) + 1 THEN GO TO ETSLOOP;
244      1    1       NOI = NO(I);  RTA = 0;  ROA = 0;  RTB = RT(NOI);  ROB = O(NOI,*);
249      1    1       IF I ^= IMT THEN DO; CALL SPLITST(NOI,NO(I+1)-2);  GO TO ETSLOOP;  END;
254      1    1       ISP = ISP - 6*(NOF-NOFGR);  CALL SPLITST(NOF,NOF);  ISP = 28;
257      1    1       IF NOF = NO(I+1)-1 THEN CALL SPLITST(NOI,NOF-1);
259      1    1                          ELSE CALL SPLITST(NOI,NOF-1);
260      1    1       ISP = 22;  CALL SPLITST(NOF+1,NO(I+1)-2);
262      1    1       ISP = 28;  PUT FILE(SPLIT) EDIT('  CMIT')(A);
264      1    1      ETSLOOP:   END TSLOOP;
                                                                          */
                    /*
                      IF COMBINED GROUP CONTAINS THE FLOATING CATEGORY, SHIFT IT TO THE
                      OTHER SIDE, AND RETRY SPLITTING
265      1         IF IMT > 0 THEN DO;  I = 0;  NOI = NO(IMT);
269      1    1       PUT FILE(SPLIT) EDIT('  CMIT')(SKIP,COL(6+NOFGR+22),A);
270      1    1       RTA = 0;  ROA = 0;  RTB = RT(NOFGR);  ROB = O(NOFGR,*);
274      1    1       CALL SPLITST(NOI,NOF-1);  ISP=ISP-6;  CALL SPLITST(NOF+1,NO(IMT+1)-2);
277      1    1                 END;
```

```
FLOSTEP:  PROC  (MATRS,HEADR,NAME,RESULT,SIGMERG,SIGSPLT,NOFL);

STMT LEVEL NEST

                  /*
                    SPECIAL CODING IF BEST SPLIT IS IN COMBINED GROUP WITH THE                */
                    FLOATING CATEGORY
  278     1       ISP = 0;  J = IM;  IPLACE = II;
  280     1       IF II = 0 THEN DO; II = IMT; IF IM > NOF THEN DO; ISP=1; IM=IM-1; END; END;
  281     1                  ELSE IF II = IMT THEN IF I4=NOF+1 THEN DO; ISP=2; IM=NOI+1; END;
  290     1                                                ELSE IF IM <= NOF THEN DO;
  296     1                                                  ISP=3; IM=IM+1; END;
  298     1     1
  301     1       SIGT = CHI(XT,DF);
  302     1       IF XT < 0 THEN DO;  PUT FILE (SPLIT) EDIT ('NO COMBINED GROUPS TO SPLIT')
  305     1     1                            (SKIP(2),A);    GO TO FINISH;    END;
  307     1       PUT FILE(SPLIT) EDIT('MAX STAT IS',XT,' TO SPLIT GROUP (',
                    SUBSTR(HEAD,NO(II),NO(II+1)-NO(II)),
                  '),  ','D.F.=',DF,',  SIGNIFICANCE=',SIGT,'%')
                    (SKIP(2),A,F(8,4),A,A,A,A,COL(64),A,F(4),A,F(9,4),A);
                  /*                                                                          */
                    IF NOT SIGNIFICANT, CONCLUDE ANALYSIS OF THIS PREDICTOR
                  IF SIGT >= SIGSPLT THEN GO TO FINISH;
  308     1       /*  SPLIT GROUP II BEFORE CATEGORY IM   */
                  O(NO(II),*) = ROT + 1.0E-30;
                  O(I4,*) =O(NO(II),*) - ROT) + 1.0E-30;  O(NO(II),*) = ROT + 1.0E-30;
  310     1       RT(IM) = RT(NO(II)) - RTT;    RT(NO(II)) = RTT;
  312     1       DO I = 1 TO M-II;  NO(M-I+2) = NO(M-I+1);  END;
  314     1       NO(II+1) = IM;   M = M + 1;
  317     1       IF II ^= IMT THEN  GO TO NLPTST;  IF IPLACE = 0 THEN NOFGR = IM;
  319     1                                                                                   */
                    SHIFT FLOATING CATEGORY IF IT IS ON WRONG SIDE
                  MATT = MAT(NOF,*);  COT = CO(NOF,*);
  323     1       IF ISP = 0 THEN GO TO ALONE;
  325     1       IF ISP >= 2 THEN GO TO SWAP2;
  327     1       SUBSTR(HEAD,NOF,IM-NOF) = SUBSTR(HEADX,NOF);  SUBSTR(HEAD,IM,1) = NOFCH;
  329     1       DO I = NOF TO IM-1;  MAT(I,*) = MAT(I+1,*);  CO(I,*) = CO(I+1,*);
  331     1                                                    NOF = IM;  GO TO ALONE;
  334     1     1   IREM(I) = IREM(I+1);  END;
                  SWAP2:  SUBSTR(HEAD,IM,NOF-IM+1)=SUBSTR(HEADX,IM-1,1);
  338     1       SUBSTR(HEAD,IM-1,1) = NOFCH;
  339     1       DO I = NOF TO IM BY -1;  MAT(I,*) = MAT(I-1,*);  END;       NOF = IM-1;
  340     1         CO(I,*) = CO(I-1,*);  IREM(I) = IREM(I-1);  END;  IREM(NOF) = NOFL;
  342     1     1 ALONE:  MAT(NOF,*) = MATT;  CO(NOF,*) = COT;  IREM(NOF) = NOFL;
  346     1       /*
                  /* CHECK WHETHER FLOATING CATEGORY IS ALONE  */
                  IF NOI ^= NOF THEN GO TO NLPTST;
  349     1       ALONE2:  IF IM ^= NOF+1 THEN GO TO NLPTST;
  351     1       NOFGR = 0;  IF IMT = 1 THEN GO TO NLPTST;
  353     1         /*
                    /* PUT IT IN POSITION 1   */
                  HEAD = NOFCH || HEADX;
  356     1       DO I = NOF TO 2 BY -1;  MAT(I,*) = MAT(I-1,*);
  357     1         CO(I,*) = CO(I-1,*);  IREM(I) = IREM(I-1);  END;
  359     1     1 MAT(1,*) = MATT;  CO(1,*) = COT;  IREM(1) = NOFL;  NOF = 1;
  362     1       DO I = IMT-1 TO 1 BY -1;  O(NO(I)+1,*) = O(NO(I),*);
  366     1         RT(NO(I)+1) = RT(NO(I));  NO(I+1) = NO(I) + 1;  END;
  368     1     1 O(1,*) = ROT;  RT(1) = RTT;   GO TO NLPTST;
  371     1
```

```
FLOSTEP:  PROC  (MATRS,HEADR,NAME,RESULT,SIGMERG,SIGSPLT,NOFL);

STMT LEVEL NEST                                                                               */

                  /*
                    ANALYSIS OF THIS PREDICTOR IS COMPLETE                                    */
  374     1       FINISH:  NOG = M;  R = 0;
                  /*
                    CALCULATE BONFERRONI MULTIPLIER
                  BONF = 1;  DO I = 1 TO M-1;  BONF = BONF*(K-1-I)/FLOAT(I);   END;           */
  376     1       /*
                    MODIFY IT IF WE HAD A FLOATING PREDICTOR
                  IF NOF ^= 0 THEN BONF = BONF * (M*(K-M-1)/FLOAT(K-2);                       */
  380     1       /*
                    DETERMINE SIGNIFICANCE OF THIS SUBTABLE
                  DO I = 1 TO M;    DO J = 1 TO L-1;   T = RT(NO(I))*MAT(K,J)/MAT(K,L);
  382     1                                                                    END;
  385     1     2   P = R + (O(NO(I),J) - T)**2 / T;   END;
                  DF = (M-1)*(L-2);  SIGMIN = CHI(R,DF);
  388     1       PUT FILE (SPLIT) EDIT('BEST IS',NOG,' GROUPS, WITH CHI-SQUARE',R,'  D.F.=',
  390     1         DF,',','BONF,' IS THE BONF MULT, AND THE SIGNIFICANCE IS',SIGMIN,'%')
                    (SKIP(2),A,F(3),A,F(9,3),A,F(4),A,F(7),A,F(14,8),A);                      */
                  /*
                    MODIFY CATEGORY NUMBERS TO CORRESPOND TO ORIGINAL NUMBERING
                  K = 0;    NO(NOG+1) = IR;
  391     1       DO I = 1 TO NOG;  K = K + 1;  NOSR(K) = NO(I+1) - NO(I);
  393     1         DO J = NO(I) TO NO(I+1)-1; K=K+1;  NOSR(K) = IREM(J);  END;  END;
  396     1     1   /*                                                                        */
                    RETURN CONTROL TO CALLING PROGRAM
                  EFINISH:  FREE CONTROL;  PUT FILE(SPLIT) EDIT((130)'*')(SKIP(2),A);  RETURN;
  401     1       /*
                    EXIT WHEN NO ANALYSIS IS PERFORMED                                        */
                  FINISHA:  NOG = 0;   SIGMIN = 100;    GO TO EFINISH;
  404     1       /*
                    SUBSECTION TO CALCULATE A SINGLE SPLIT STATISTIC
                                                  DO J = II TO I2;
                  SPLITST:  PROC (II,I2);          ROB = ROB - CO(J,*);
  407     1         ROA = ROA + CO(J,*);
  409     2     1   RTA = RTA + MAT(J,IC);        RTB = RTB - MAT(J,IC);
  411     2     1   FEA = O(NO(I,*)*RTA/RT(NO(I));  REB = O(NO(I,*)*RTB/RT(NO(I));
  413     2     1   R = SUM((ROA-REA)**2/REA + (ROB-REB)**2/REB);
  415     2     1   PUT FILE (SPLIT) EDIT (R) (COL(6*J+ISP),F(6,1));  /* FIX UP SPACING  */
  416     2     1   IF R > XT THEN DO; XT=R;  IM=J+1;  II=I;  RTT=RTA;  ROT=ROA;  END;
  417     2     1     END SPLITST;
  425     2     1                                  END FLOSTEP;
  427     1
```

```
FRESTEP:  PROC (MATRS,HEADR,NAME,RESULT,SIGMERG,SIGSPLT);

STMT LEVEL NEST
  1                FRESTEP:  PROC (MATRS,HEADR,NAME,RESULT,SIGMERG,SIGSPLT);
                     /*
                       THIS PROCEDURE ANALYSES IN DETAIL A FREE PREDICTOR ON THE BASIS OF      */
                       A CONTINGENCY TABLE PROVIDED BY THE CALLING PROGRAM
  2     1          DCL  1 MATRS, 2 MATR(*,*),  HEADR CHAR(*) VAR,    SPLIT PRINT,
                     NAME CHAR(*) VAR,  TITLE CHAR(50) VAR EXT,   PHASE CHAR(11),
                     1 RESULT, 2 SIGMIN, 2 RONE, 2 NOG, 2 NOS(32),    NOTE(32),
                     NO(32) INIT (1,1,1,2,1,3,1,4,1,5,1,6,1,7,1,8,1,9,1,10,1,11,1,12,1,13,
                       1,14,1,15,1,16),
                     IW(17) INIT (1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17),
                     1 CONTROL CTL, 2 MAT(K,L), 2 (E,O,CE,CO) (K-1,L-1), 2 (RT,IREM)(K),
                       2 (REA,ROA,KEB,ROB,RET,ROT)(L-1),  2 HEAD CHAR(8);
  3     1          DCL  BITS(16) BIT(1);                                                    */
                     /*
                       PROVIDE HEADINGS FOR OUTPUT
  4     1          ON ENDPAGE(SPLIT) PUT FILE(SPLIT) EDIT('G V KASS. 1974. V.A.',TITLE,
                     'CONTINUED',**) (A,SKIP,A,SKIP,COL(100),A,SKIP,A) PAGE;
  6     1          IR = DIM(MATR,1);   IC = DIM(MATR,2);   K, L = 1;
                     /*                                                                     */
                       REMOVE ROWS AND COLUMNS WITH ZERO MARGINAL TOTALS
  9     1          DO I = 1 TO IR-1;  IF MATR(I,IC) > 0 THEN  K = K + 1;  END;
 13     1          DO I = 1 TO IC-1;  IF MATR(IR,I) > 0 THEN  L = L + 1;  END;
 17     1          ALLOCATE CONTROL;  IS = 0;  BONE = 1;
 20     1          DO I = 1 TO IR;  IF MATR(I,IC) <= 0 THEN GO TO END1;
 23     1  1         IS=IS+1;  IREM(IS)=I;  SUBSTR(HEADR,IS,1) = SUBSTR(HEADR,I,1);  IT=0;
 27     1  1         DO J = 1 TO IC;  IF MATR(IR,J) <= 0 THEN GO TO END2;
 30     1  2           IT=IT+1;  MAT(IS,IT) = MATR(I,J);
 32     1  2         END2:  END;        END1:  END;              M = K -1;
 36     1          IF LINENO (SPLIT) > 53 - M THEN SIGNAL ENDPAGE (SPLIT);
 37     1          PUT FILE (SPLIT) EDIT((130)'*','FREE PREDICTOR:  ',NAME)(SKIP,A,SKIP,A,A);
                                                                                           */
                     /*
                       SUSPEND ANALYSIS IF LESS THAN TWO ROWS OR COLUMNS
 38     1          IF M<=1 THEN DO; PUT FILE (SPLIT) EDIT (' IS A CONSTANT.')(A);
 41     1  1           GO TO FINISHA;  END;
 43     1          IF L<=2 THEN DO; PUT FILE (SPLIT) EDIT('DEPENDENT VARIABLE IS A CONSTANT.')
 46     1  1         (SKIP(2),A);  GO TO FINISHA;  END;                                    */
                     /*
                       CALCULATE EXPECTED VALUES
 48     1          X = 0;   DO I = 1 TO M;  R = 0;   DO J = 1 TO L-1;
 52     1  2         E(I,J)=FLOAT(MAT(I,L)*MAT(K,J))/MAT(K,L);  O(I,J)=MAT(I,J);
 54     1  2         IF O(I,J) <= 1.0E-30 THEN O(I,J) = 1.0E-30;
 56     1  2         P = R + (O(I,J)-E(I,J))**2 / E(I,J);   END;      X = X + R;   END;
 60     1          DF = (M-1)*(L-2);  SIG = CHI(X,DF);  CE = E;   CO = O;
 64     1          PUT FILE(SPLIT) EDIT('TABLE HAS CHI-SQUARE',X,' WITH D.F.',DF,
                     ', AND SIGNIF',SIG,'%') (SKIP,A,F(9,3),A,F(4),A,F(13,8),A);
 65     1          DF = L-2;  NLOOP = 0;    MAXLOOP =3*M;                               */
                     /*
                       THIS LOOP MERGES AND SPLITS THE CATEGORIES
 71     1          LOOP:  PHASE = 'MERGE PHASE';
 72     1            IF M <= 1 THEN GO TO TRYSPLT;
```

```
FRESTEP:  PROC (MATRS,HEADR,NAME,RESULT,SIGMERG,SIGSPLT);

STMT LEVEL NEST                                                                            */
                     /*
                       THIS SUBSECTION PRINTS A SUBHEADING          HEADING:      PROC;
 74     1          CALL HEADING;
 76     2          IF LINENO(SPLIT) > 56-M THEN SIGNAL ENDPAGE (SPLIT);
 76     2          PUT FILE(SPLIT) EDIT(PHASE,M,' GROUPS:',' ')(SKIP(2),A,F(5),A,COL(27),A);
 79     2          IP = 1;
 80     2          DO I = 2 TO M; PUT FILE(SPLIT) EDIT(SUBSTR(HEAD,NO(IP+1),1))(A,A);     END;
 82     2  1         DO J = 1 TO NO(IP)-1;
 83     2  2           PUT FILE(SPLIT) EDIT(',',SUBSTR(HEAD,NO(IP+J+1),1))(A,A);
 85     2  1         IP = IP + NO(IP) + 1;      END;                                       END HEADING;
                     /*                                                                     */
                       CALCULATE AND PRINT TEST STATISTICS FOR MERGING
                     IP = NO(1) + 2;   II = 0;    XMIN = 1.E60;
 88     1          CAP:  DO I = 2 TO M;
 91     1            PUT FILE(SPLIT) EDIT(SUBSTR(HEAD,NO(IP+1),1))(SKIP,COL(30-2*NO(IP)),A);
 92     1  1         DO J = 1 TO NO(IP)-1;
 93     1  1           PUT FILE(SPLIT) EDIT(',',SUBSTR(HEAD,NO(IP+J+1),1))(A,A);     END;
 94     1  2           IPJ = 1;
 96     1  1         IP = IP + NO(IP) + 1;
                     /*
                       /*  EXAMINE ROWS  I  &  J      */
 98     1  1         DO J = 1 TO I-1;
 99     1  2           ROT = (O(IW(I),*) + O(IW(J),*)) / (RT(IW(I)) + RT(IW(J)));
100     1  2           REA = ROT*RT(IW(I));      REB = ROT*RT(IW(J));  REB=1-REB;  REB=2/REB;
102     1  2           P = SUM((O(IW(I),*)-REA)**2/REA + (O(IW(J),*)-REB)**2/REB);   IPJ=IPJ+NO(IPJ)+1;
103     1  2           PUT FILE(SPLIT) EDIT(R)(X(2*NO(IPJ)-2),F(6,1));                    */
                     /*
                       FIND MINIMUM MERGE STATISTIC
                       IF R < XMIN THEN DO; XMIN = R;  II = I;  JJ = J;  END;
105     1  2       END CAP;
111     1  2       IF II = 0 THEN DO; PUT FILE(SPLIT) EDIT('CANNOT EXPLAIN ANYTHING')
113     1            (SKIP,A);  GO TO FINISH;   END;
116     1  1       SIGS = CHI(XMIN,DF);
118     1          PUT FILE(SPLIT) EDIT('MIN STAT IS',XMIN,',  TO MERGE GROUPS NUMBERED',
119     1            JJ,' AND',II,'.  ','D.F.=',DF,',  SIGNIFICANCE',SIGS,'%')
                     (SKIP(2),A,F(8,4),A,F(3),A,F(3),A,COL(64),A,F(4),A,F(9,4),A);      */
                     /*
                       IF CANNOT MERGE TRY SPLITTING
                     IF SIGS < SIGMERG THEN GO TO TRYSPLT;
120     1          I = II;   II = JJ;   JJ = I;
122     1          /*                                                                     */
                     /* COMBINE GROUPS IN POSITIONS II & JJ     */
                     O(IW(II),*)=O(IW(II),*)+O(IW(JJ),*);  RT(IW(II))=RT(IW(II))+RT(IW(JJ));
125     1          M1 = M-1;   DO I = JJ TO M1;  IW(I) = IW(I+1);  END;                */
127     1            /*
                       RESET POINTERS
                     SETMAX:   /* REDO NUMBER   */
131     1            IP=1;     DO I = 1 TO II-1;  IP=IP+NO(IP)+1;  END;
132     1            IPJ=IP;   DO I = II TO JJ-1;  IPJ=IPJ+NO(IPJ)+1;  END;
135     1            IP1 = IP + NO(IP) + 1;    IPJ1 = IPJ + NO(IPJ) + 1;
139     1            DO I = IP1 TO IPJ1-1;  NOTE(I) = NO(I);  END;
141     1            NO(IP) = NO(IP) + NO(IPJ);   IA = 0;
144     1            DO I=IPJ+1 TO IPJ1-1;  NO(IP1+IA)=NOTE(I);  IA=IA+1;  END;
146     1            DO I=IP1 TO IPJ-1;  NO(IP1+IA)=NOTE(I);  IA=IA+1;  END;      M = M1;
150     1            DO I = IPJ1-1 TO DIM(NO,1)-1;  NO(I)=NO(I+1);  END;
154     1
```

```
         FRESTEP:  PROC (MATRS,HEADR,NAME,RESULT,SIGMERG,SIGSPLT);

STMT LEVEL NEST

                     /*
                       TEST IF PROCEDURE IS IN A LOOP --  EMERGENCY EXIT                */
 158     1          *LPTST:  NLOOP=NLOOP+1;
 159     1           IF NLOOP > MAXLOOP THEN DO;  PUT EDIT('FORCE TERMINATION - IN A LOOP')
 162     1     1       (SKIP(2),A);   PUT FILE(SPLIT) EDIT('FORCE TERMINATION - IN A LOOP')
 163     1     1           (SKIP(2),A);  GO TO FINISH;  END;
 165     1           GO TO LOOP;
                     /*
                       TRY SPLITTING A COMBINED GROUP                                   */
 166     1          TRYSPLT:  PHASE='SPLIT PHASE';
                     /*
                       NO ATTEMPT IS MADE TO SPLIT IF LEVEL OF MERGING IS LESS THAN THAT
                       OF SPLITTING                                                     */
 167     1           IF SIGMERG <= SIGSPLT THEN GO TO FINISH;
 169     1           IF LINENO (SPLIT) > 55 THEN SIGNAL ENDPAGE (SPLIT);
                     /*
                       PROVIDE HEADING                                                  */
 171     1          PUT FILE(SPLIT) EDIT(PHASE,M,' GROUPS:',' ')(SKIP(2),A,F(5),A,COL(29),A);
 172     1           IP = 1;  DO I = 1 TO M;
 174     1     1       PUT FILE(SPLIT) EDIT(' ( ',SUBSTR(HEAD,NO(IP+1),1))(A,A);
 175     1     1       DO J = 2 TO NO(IP);                                        END;
 176     1     2         PUT FILE(SPLIT) EDIT(SUBSTR(HEAD,NO(IP+J),1))(X(5),A);
 178     1     1       PUT FILE(SPLIT) EDIT(' )')(A);  IP = IP + NO(IP) + 1;    END;
 181     1           IP = 1;  XT = -1;
                     /*
                       EXAMINE GROUPS WITH MORE THAN ONE CATEGORY                       */
 183     1          DO I = 1 TO M;  IF NO(IP) <= 1 THEN GO TO NEXTGR;
 185     1     1       NOIP = NO(IP);  IWHICH = 0;  NOCOL = 1;  IFIRST = 17 - NOIP;
 190     1     1       PUT FILE(SPLIT) EDIT('TEST STATS FOR SPLITTING THE GROUP:  ')(SKIP(2),A);
 191     1     1       DO J = 1 TO NOIP;
 192     1     2         PUT FILE(SPLIT) EDIT(SUBSTR(HEAD,NO(IP+J),1))(X(1),A);    END;
                       /*
                         /*  FIND WHICH GROUPS TO FORM.  */
 194     1     1       FWGTF:  IWHICH = IWHICH + 1;
 195     1     1       STRING(BITS) = UNSPEC(IWHICH);  IF BITS(IFIRST) THEN GO TO NEXTGR;
 198     1     1      PUT FILE(SPLIT) EDIT((BITS(J) DO J=IFIRST TO 16))(COL(NOCOL),16 A(1));
 199     1     1       NOCOL = NOCOL + NOIP + 8;  IF NOCOL > 124 - NOIP THEN NOCOL = 1;
 202     1     1       RTA = 0;  ROA = 0;  ROB = 0;  RTB = 0;
                       /*
                         CALCULATE SPLIT STATISTIC FOR SPLIT SPECIFIED BY "BITS"        */
 206     1     1       DO J = 1 TO NOIP;  IF BITS(IFIRST+J-1)
 208     1     2       THEN DO; RTB=RTB+MAT(NO(IP+J),L); ROB=ROB+CO(NO(IP+J),*);  END;    END;
 212     1     2       ELSE DO; RTA=RTA+MAT(NO(IP+J),L); ROA=ROA+CO(NO(IP+J),*);  END;
 217     1     1       REA = O(IW(I),*)*RTA/RT(IW(I));  REB = O(IW(I),*)*RTB/RT(IW(I));
 219     1     1       R = SUM((ROA-REA)**2/REA + (ROB-REB)**2/REB);
 220     1     1       PUT FILE (SPLIT) EDIT (R) (F(6,1));
```

```
         FRESTEP:  PROC (MATRS,HEADR,NAME,RESULT,SIGMERG,SIGSPLT);

STMT LEVEL NEST

                                                                                      */
                     /*
                       FIND MAXIMUM SPLIT STATISTIC
 221     1     1       IF R > XT THEN DO; XT=R; II=I; JJ=IWHICH; IPS=IP; RTT=RTA; ROT=ROA; END;
 230     1     1       GO TO FWGTF;
 231     1     1      NEXTGR:  IP = IP + NO(IP) + 1;              END;
                                                                                      */
                     /*
                       EXAMINE THE BEST SPLIT FOUND
 233     1          SIGS = CHI(XT,DF);  IFIRST = 17-NO(IPS);  IWHICH = JJ;
 236     1          IF XT < 0 THEN DO;  PUT FILE (SPLIT) EDIT ('NO COMBINED GROUPS TO SPLIT')
 239     1     1                              (SKIP(2),A);     GO TO FINISH;     END;
 241     1          PUT FILE(SPLIT) EDIT('MAX STAT IS',XT,' TO SPLIT GROUP NUMBERED';II,
                      ' AS ',SUBSTR(UNSPEC(JJ),IFIRST,17-IFIRST),
                      ',  ','D.F.:',DF,',  SIGNIFICANCE',SIGS,'%')
                       (SKIP(2),A,F(8,4),A,F(3),A,A,A,COL(64),A,F(4),A,F(9,4),A);
                                                                                      */
                     /*
                       CONCLUDE ANALYSIS IF BEST SPLIT NOT SIGNIFICANT ENOUGH
 242     1          IF SIGS > SIGSPLT THEN GO TO FINISH;
                     /*
                     /* IMPLEMENT THIS SPLIT   */
 244     1          JJ =  INDEX(UNSPEC(IWHICH),'1'B)-IFIRST;  NOIP = NO(IPS);
 246     1          DO I = 1 TO M-II;  IW(M-I+2) = IW(M-I+1);  END;
 249     1          IW(II+1) = NO(IPS+JJ+1);
 250     1          RT(IW(II+1)) = RT(IW(III)) - RTT;   RT(IW(III)) = RTT;
 252     1          O(IW(II+1),*) =(O(IW(III),*) - ROT) + 1.0E-30;  O(IW(III),*) = ROT + 1.0E-30;
                     /* REDO THE  NO(*) ARRAY   */
 254     1          DO I = 1 TO 31-IPS-NOIP;  NO(33-I) = NO(32-I);  END;
 257     1          STRING(BITS) = UNSPEC(IWHICH);  NOTE = NO;  J = 0;  JJ = 0;
 261     1          DO I = 1 TO NOIP;  IF BITS(IFIRST+I-1)
 263     1     1       THEN DO;  JJ = JJ + 1;  NO(IPS+NOIP-JJ+2) = NOTE(IPS+I);  END;     END;
 267     1     1       ELSE DO;  J = J + 1;  NO(IPS+J) = NOTE(IPS+I);  END;  GO TO NLPTST;
 272     1          NO(IPS) = JJ;  NO(IPS+J+1) = NOIP-JJ;  M = M + 1;
```

```
        FRESTEP:  PROC (MATRS,HEADR,NAME,RESULT,SIGMERG,SIGSPLT);

STMT LEVEL NEST

                    /*                                                                    */
                      CONCLUDE ANALYSIS
276     1         FINISH:  NOG = M;  R = 0;
                    /*                                                                    */
                      CALCULATE BONFERRONI MULTIPLIER
278     1          IF M <= 1 THEN BONF = 1;  ELSE DO;  FM = 1;  BONF = 0;  FI = 1;  K1 = K-1;
285     1    1       DO I = 2 TO M;  FM = FM*I;  END;       FM = 1/FM;
289     1    1       DO I = 1 TO M;  FI = -FI*I;  FM = FM*(M-I+1);
297     1    2       BONF = BONF + FM*(FLOAT(I)**K1)/FI;  END;          BONF = ABS(BONF);   END;
                    /*                                                                    */
                      DETERMINE SIGNIFICANCE OF THIS SUBTABLE
296     1          DO I = 1 TO M;    DO J = 1 TO L-1;
298     1    2       T = RT(I)*(I)) = MAT(K,J) / MAT(K,L);
299     1    2       R = R + (OE(I)*(I),J) - T)**2 / T;     END;              END;
302     1          DF = (M-1)*(L-2);  SIGMIN = CHI(R,DF);
304     1          PUT FILE (SPLIT) EDIT('BEST IS',NOG,' GROUPS, WITH CHI-SQUARE',R,'  D.F.=',
                     DF,',',BONF,' IS THE BONF MULT, AND THE SIGNIFICANCE IS',SIGMIN,'%')
                     (SKIP(2),A,F(3),A,F(9,3),A,F(4),A,F(7),A,F(14,8),A);
305     1          IP = 1;     DO I = 1 TO NOG;     DO J = 1 TO NO(IP);
308     1    2                    NOS(IP+J) = IREM(NO(IP+J));     END;
310     1    1       NOS(IP) = NO(IP);    IP = IP + NO(IP) + 1;   END;
                                                                                        */
                    /*
                      RETURN CONTROL TO CALLING PROGRAM
313     1         FINISHE:  FREE CONTROL;  PUT FILE(SPLIT) EDIT((130)'*')(SKIP(2),A);  RETURN;
                                                                                        */
                    /*
                      EXIT WHEN NO ANALYSIS PERFORMED
316     1         FINISHA:  NOG = 0;  SIGMIN = 100;  GO TO FINISHE;       END FRESTEP;
```

```
        CHI:  PROC (Y,DF) RETURNS (DEC FLOAT(6));

STMT LEVEL NEST
  1             CHI:  PROC (Y,DF) RETURNS (DEC FLOAT(6));
                    /*                                                                    */
                      THIS PROCEDURE CALCULATES THE SIGNIFICANCE OF A GIVEN CHI-SQUARE
                      VARIATE
                    /*                                                                    */
  2     1         DCL (Z,P,F) STATIC,  RTH INIT (.70710678) STATIC;
                    /*                                                                    */
                      SPECIAL CASES OF OBVIOUS NON SIGNIFICANCE
  3     1          IF Y <= 0 THEN GO TO NOTSIG;     IF DF <= 0 THEN GO TO NOTSIG;
                    /*                                                                    */
                      USE SQUARE OF NORMAL DISTRIBUTION FOR 1 DEGREE OF FREEDOM
  7     1          IF DF = 1.E0 THEN DO;  P = 100*ERFC(RTH*SQRT(Y));  RETURN (P);  END;
                    /*                                                                    */
                      TWO DEGREES OF FREEDOM IS EXPONENTIAL DISTRIBUTION
 12     1          IF DF = 2.E0 THEN DO;  P = 100*EXP(-Y/2);  RETURN (P);  END;
                    /*                                                                    */
                      USE WILSON-HILFERTY APPROXIMATION                    */
                    /* FROM JASPEN No. (1964)  EDUC & PSYCHO MEAS., 25, 877-880.
 17     1          IF Y > DF THEN F = Y/DF;  ELSE  F = DF/Y;
 20     1          Z = (F**.3333333 - 1 + .2222222/DF) / SQRT(.2222222/DF);
 21     1          P = 50*ERFC(RTH*Z);     IF Y > DF THEN RETURN (P);  ELSE  RETURN (100-P);
 25     1         NOTSIG:   P= 100;   RETURN (P);     END CHI;
```

```
PRINTAB: PROC (MATS,HEAD,NAME,RESULT,DEPNAMES,P,SIG,TAB);

STMT LEVEL NEST
  1                 PRINTAB: PROC (MATS,HEAD,NAME,RESULT,DEPNAMES,P,SIG,TAB);
                      /*
                        THIS ROUTINE PRINTS THE GIVEN CONTINGENCY TABLE, MERGING
                        CATEGORIES AS SPECIFIED                                    */
  2     1             DCL  1 MATS, 2 MAT(*,*),   HEAD CHAR(*) VAR,
                      NAME CHAR(*) VAR,  DEPNAMES(*) CHAR(*) VAR,
                      TITLE CHAR(500) VAR EXT,
                      1 RESULT, 2 SIGMIN, 2 BONF, 2 NOG, 2 NOS(32),    ID, P CHAR(1), D(NOG+1),
                      I, J, K, L, N, IP, (TOT,RTOT) BIN FIXED, TAB PRINT;
                      /*                                                           */
                        PREPARE PAGE HEADING
  3     1             ON ENDPAGE(TAB) PUT FILE(TAB)EDIT('G V KASS. 1974. V.A.',TITLE,'CONTINUED')
                                    (A,SKIP,A,SKIP,COLUMN(100),A) PAGE;
  5     1             K = DIM(MAT,1);  L = DIM(MAT,2);
  7     1             IF LINENO(TAB) > 56-L THEN SIGNAL ENDPAGE (TAB);
  9     1             PUT FILE(TAB) EDIT((130)'*','PREDICTOR:  ',NAME)(SKIP(2),A,SKIP,A,A);
                      /*
                        DETERMINE WHETHER OUTPUT SHOULD PROVIDE -
                          1.  NUMBERS PER CELL
                          2.  ROW PERCENTAGES
                          3.  OVERALL PERCENTAGES                                  */
 10     1             IF P = 'P' THEN DO;  ID = 1;  IP = 1;  D(NOG+1) = 100.E0/MAT(K,L);
 15     1     1         PUT FILE (TAB) EDIT (' = ROW PERCENTAGES *') (A);
 16     1     1         DO J = 1 TO NOG;  TOT = 0;
 18     1     2           DO N = 1 TO NOS(IP);  TOT = TOT + MAT(NCS(IP+N),L);  END;
 21     1     2           IF TOT = 0 THEN D(J) = 0;  ELSE D(J) = 100.E0/TOT;
 24     1     2           IP = IP + NOS(IP) + 1;   END;                    END;
 27     1               ELSE
 27     1               IF P='P' THEN DO;  ID = 0;  D = 1;  END;
 32     1                     ELSE DO; ID=1; D=1.E0/MAT(K,L);
 35     1     1                 PUT FILE(TAB)EDIT(' *PERCENTAGES*  TOTAL NUMBER',MAT(K,L))
 36     1     1                                 (A,F(7));                END;
                      /*                                                           */
                        PRINT HEADING TO TABLE
 37     1             PUT FILE (TAB) EDIT (' ') (SKIP,COLUMN(21),A);
 38     1             IP=1;  DO I=1 TO NOG;
 40     1     1       PUT FILE (TAB) EDIT (SUBSTR(HEAD,NOS(IP+1),1)) (X(5),A);
 41     1     1         DO J=1 TO NOS(IP)-1;
 42     1     2           PUT FILE (TAB) EDIT (',',SUBSTR(HEAD,NOS(IP+1+J),1)) (A,A);   END;
 44     1     1         IP = IP + NOS(IP) + 1;    END;
 46     1             PUT FILE (TAB) EDIT ('TOTAL') (X(5),A);
```

```
PRINTAB: PROC (MATS,HEAD,NAME,RESULT,DEPNAMES,P,SIG,TAB);

STMT LEVEL NEST
                      /*                                                           */
                        CALCULATE CELL VALUES ROW BY ROW
 47     1             DO I = 1 TO L;  IP = 1;  RTOT = 0;
                      /*                                                           */
                        CHECK IF LAST ROW  -TOTALS-
 50     1     1       IF I=L
 51     1     1       THEN PUT FILE(TAB)EDIT(DEPNAMES(I))(SKIP,COLUMN(22-LENGTH(DEPNAMES(I))),A);
 52     1     1       ELSE   IF P = 'R' THEN DO;  ID = 0;  D = 1;
 56     1     2                 PUT FILE (TAB) EDIT ('TOTALS (100%)') (SKIP,COL(8),A);  END;
 58     1     1            ELSE PUT FILE (TAB) EDIT ('TOTALS') (SKIP,COLUMN(15),A);
 59     1     1       DO J=1 TO NOS; TOT=0; DO N =1 TO NOS(IP); TOT=TOT+MAT(NOS(IP+N),L); END;
 64     1     2       PUT FILE (TAB) EDIT (TOT*D(J)) (X(2*NOS(IP)-2),F(6,ID));
 65     1     2       RTOT = RTOT + TOT;   IP = IP + NOS(IP) + 1;   END;
 68     1     1       PUT FILE (TAB) EDIT (RTOT*D(NOG+1)) (X(4),F(6,ID));
                      /*                                                           */
                        PRINT SIGNIFICANCE LEVEL OF TABLE
 70     1             PUT FILE(TAB) EDIT('TABLE IS SIGNIFICANT AT',SIG,'% LEVEL')
                        (SKIP,COL(40),A,F(10,5),A);
 71     1             RETURN;          END PRINTAB;
```

```
TABSIG:  PROC (MATS,X,DF,SIG);

STMT LEVEL NEST
  1                 TABSIG:  PROC (MATS,X,DF,SIG);
                      /*
                        THIS SECTION CALCULATES THE SIGNIFICANCE OF A GIVEN CONTINGENCY      */
                        TABLE.   NO GROUPING OF CATEGORIES PERFORMED.
  2     1           DCL  1 MATS,  2 MAT(*,*);
  3     1           K = DIM(MAT,1);  L = DIM(MAT,2);  SIG = 0;  DF = 0;  X = 0;
  4     1           IF MAT(K,L) <= 1 THEN RETURN;     NI = 0;  NJ = 0;
                      /*
                        SUM STATISTIC CELL BY CELL, IGNORING ROWS AND COLUMNS WITH ZERO      */
                        MARGINAL TOTALS
 12     1           DO I = 1 TO K-1;   IF MAT(I,L) <= 0 THEN GO TO ELI;
 15     1     1      NI = NI + 1;   DO J = 1 TO L-1;   IF MAT(K,J) <= 0 THEN GO TO ELJ;
 18     1     2        NJ = NJ + 1;  E = FLOAT(MAT(I,L)*MAT(K,J)) / MAT(K,L);
 21     1     2        X = X + (MAT(I,J)-E)**2 / E;
 22     1     2     ELJ:  END;     ELI:  END;                                              */
                      /*
                        CALCULATE DEGREES OF FREEDOM AND SIGNIFICANCE
 24     1           DF = NJ - NI + 1 - NJ/NI;   SIG = CHI(X,DF);
 26     1           RETURN;          END TABSIG;
```

Appendix C.

This appendix concerns the analysis of 669 CTA students that is discussed in chapter 9.

The following pages give a listing of the data and some of the output of the program with circled numbered pages for ease of reference in the text.

The data for each student is coded into eight contiguous printed columns where

Col 1 - Year of entry
Col 2 - Type of matriculation (dependent variable)
Col 3 - Whether eventually qualified or not
Col 4 - Number of commercial subjects
Col 5 - Number of non-academic subjects
Col 6 - Type of school
Col 7 - Matriculation Mathematics mark
Col 8 - Matriculation English mark.

LISTING OF 669 CTA STUDENTS.

```
21321354 21311265 21311244 11311255 21441243 11431343 21311275 11311264 21441344 11211253 21311245 11311264
21431334 11321264 21321245 21311265 11321265 21311244 11331264 21311255 11311245 21211253 11211253 21311265
21322264 21211244 21311244 21311255 21211223 21312264 11311255 21211334 21121243 21321266 21211234 21321245
21312244 21311275 21321245 21221243 21321244 11321255 11311265 21322274 11321265 21311254 21421233 11211265
21111265 11322254 21431335 21431337 11121244 11311255 21321265 11311265 11311245 21211333 21311264 21311244
21321254 21321275 21211233 11211265 11421333 11311265 21321244 21312254 21321245 21211243 21321255 21311254
21311244 21411355 11111264 21321245 11311256 21411244 21311254 21321254 11121245 21211256 21211233 21321244
21421333 11311265 21111244 21312254 21311245 21311254 21311264 11311264 21321245 21311254 21311244 21211234
21321265 21311254 21321244 21211255 21311233 21311245 11311264 21311255 11211255 21311266 11421354 21431334
11211254 21312264 21211255 21211243 21311266 11311275 21321276 11211233 11111256 21321244 11311244 21322264
11211233 11211245 21211233 21211233 11311254 11311275 21421344 11211244 21211233 21311254 21211244 11322244
21321255 11211234 11321264 11311266 11211256 21312244 21321244 11311276 11311267 21211245 21211234 11211234
21211233 21211233 21311266 21321244 11311265 21211233 21321244 21331245 21421244 11321265 21311276 11311245
21441345 21211253 11311265 21311245 21311255 21311244 12311245 12311244 12311233 22431343 21312245 11311265
11221253 11311264 22311254 12311254 22441354 12311255 12311244 12311244 22211233 22312211 22321244 12321254
22211243 22441334 22311254 12312275 22321244 12321254 22211263 12322244 12311244 12211263 12211243 12322254
22441363 22211243 22311255 12322254 22111233 12421233 12322264 22311264 22451233 22311244 22311263 12312243
22311255 22311255 22441343 12311254 12311244 12421354 12311273 12311233 22331233 12311244 22321243 22311235
22321244 22121233 12321254 12311253 12311255 12311254 22211234 12311275 12311233 22211345 12121254 22311245
12311264 12311274 12311244 22211243 22111243 22111234 22311243 22311243 22331244 22311245 22211235 22311244
22311245 22321234 22331254 22311276 22211233 22311254 12311253 12211243 22232244 12311245 22321233 12431353
22311244 12311265 12311275 12321255 22321235 22321243 22311255 22311354 12311275 12311243 22312244 22212211
22211253 12311274 22312245 12311244 22221265 22321254 12211265 22211243 22311243 22211264 22211244 22312266
22421333 22311244 12441333 12211244 12321254 12211235 22121254 12221243 12311264 22311254 22211244 12311256
12211275 12431343 22312274 12311266 22311243 12211244 22211233 22221211 22311254 22311254 12321274 12311254
12311254 22311233 12311276 22112244 12321253 12322234 12322273 12311234 12311255 22441344 22211236 12311245
22311243 12321264 12321254 22421353 22121265 12322234 22211243 22311233 12311243 12321244 13311253 13211234
22311245 22311253 12322244 22111247 22211243 22311244 12211243 22321244 12321234 12211266 23421344 13321243
22421233 22112233 12441364 12321254 22111263 12311254 12211243 13321253 13121244 23511234 23321244 13311243
13121244 13322233 23441333 13311274 13311264 23311244 23321244 23311244 13311265 13441363 23441344 13311254
13431333 13391254 13311253 13312253 23221343 23321243 13221244 13311255 13311253 13211343 13211255 23311244
13311244 23111234 13211234 13211245 23311243 23311255 13211254 23311254 23311254 13321254 13311254 23451334
23311253 13211255 23321253 13311264 23211343 23211345 23311254 23311243 23311243 13121264 23311233 13311254
23311244 23311243 23211234 23321254 13431343 13411254 23311254 23311254 23411253 23211244 23311244 13511235
23321263 13321233 23311233 23311233 23321233 23221253 23441333 23322233 23311254 13311253 23311275 23441364
13311233 13311253 23311233 13121263 23311235 23311243 23211254 13211234 13311244 13311234 13431333 13311254
23321233 23322243 23312263 23311243 23311255 13321274 23521233 13211244 13431334 13121264 23421345 13211255
13311233 23311233 23212233 13321244 23311233 23511234 13311256 23311233 13431256 13121264 23311234 13321256
23321233 23311233 23431335 23311244 23111234 23441334 13211243 13211243 13311253 23211264 23311234 13321254
23441333 13321243 13421244 13211275 23441355 13311244 13411233 13211263 13211243 23311233 23441243 23311234
13321234 13311234 23321244 23321256 23321234 23522243 23211233 23441334 13311254 23311254 14522243 13321253
13311263 23211244 13511244 13511264 23421333 13212255 13211262 23441334 24521235 24431333 14511233 14522243
13311243 13511243 14511253 14511243 24451353 14521253 24421354 24441364 14511275 24511244 14211243 14511243
23441333 13311263 24511243 24112244 14521233 24211267 24421343 14111264 14211255 24512233 24431333 24431333
24111244 24511243 24519253 14511234 14511255 14511254 24511244 24521245 24511275 24431333 14211233 24431333
14211254 24511235 24521234 24211263 24441233 24451354 24511243 14511243 14311243 24511275 24512234 24431334
14311243 24511254 24211274 14411355 24531233 24211265 24211233 14322264 14511234 24221343 24512253 14512244
14521243 24311254 24451353 24211264 24441353 24441364 14441336 24441353 24511234 24211244 24511233 24141344
14511244 24441334 24511273 14211256 14511273 24121253 24511234 14311244 24211353 24211353 24511253 24312244
14511234 24511234 24111235 14511233 24511243 24221234 14411253 24511243 14511233 24511244 24511234 14211244
24521244 24321243 14511253 14121244 24211233 24211253 24111211 14511254 14511254 14512265 14431355 24441354
24141344 24441343 24441343 14121244 24511233 24211253 14211243 24511254 24441343 14511275 14431353 24211344
14211244 14211255 24221233 24441354 24441354 24441344 24511243 14511254 14511244 14512234 24221253 24211362
24212233 14211244 14522234 14311234 14441343 14511243 24441354 24211244 24211244 24311234
24511233 24511254 24121275 24121244 14121245 14511233 14511244
24511244 24451373 14511253 24211234 14521244 14311254 14211263
24521244 24211333 24211233 14431243 14521235 14511254
```

(1)

WHICH MATRIC DID CTA STUDENT WRITE? HAS 4 CATEGORIES CALLED:
OTHER
JKB
TSSC & TUEC
NSC

THERE ARE 7 PREDICTORS AS FOLLOWS

TYPE	NO. OF CATEGORIES	CATEGORY SYMBOLS	USE?	MERGE%	SPLIT%	NAME
MONO	2	YN	NOT	1.00	0.95	DID STUDENT EVENTUALLY GET THE CTA (YES,NO)
FREE	4	0123	MAY	1.00	0.95	YEAR OF ENTRY (60, 61, 62, 63)
FREE	5	01234	MAY	1.00	0.95	NUMBER OF COMMERCIAL SUBJECTS
FREE	2	01	MAY	1.00	0.95	NUMBER OF NON-ACADEMIC SUBJECTS
FREE	3	ABC	MAY	1.00	0.95	TYPE OF SCHOOL
FLOAT 1	7	7FEDCBA	MAY	1.00	0.95	MATHEMATICS MARK
FLOAT 1	7	7FEDCBA	MAY	1.00	0.95	ENGLISH MARK

OPTION 1 IS 0
OPTION 2 IS 0
OPTION 3 IS 100
OPTION 4 IS 1
OPTION 5 IS 1
OPTION 6 IS 20

OPTIONS IN EFFECT:
TABLES PRINTED WITH ROW PERCENTAGES
TABLES WILL BE PRINTED BEFORE & AFTER EVERY STEP
A GROUP WITH LESS THAN 100 CASES WILL NOT BE ANALYSED
A GROUP WILL BE SPLIT IF IT IS SIGNIFICANT AT THE 1% LEVEL
A GROUP WILL BE SPLIT IF IT IS BONFERRONI SIGNIFICANT AT THE 1% LEVEL
THE PROGRAM WILL TERMINATE WHEN 20 OR MORE GROUPS HAVE BEEN FORMED OR ANALYSED

G V KASS. 1974

END OF ANALYSIS

THIS IS THE VERY END OF THE ANALYSIS. G V KASS 1974

(2)

SUMMARY OF RESULTS
ANALYSIS OF TOTAL GROUP.

NO.	NAME	SIGNIF (%)	BONF-SIG (%)	NUMBER OF GROUPS.	
1	DID STUDENT EVENTUALLY GET THE CTA (YES,NO)	0.46708707	0.48708707	2	Y N
2	YEAR OF ENTRY (60, 61, 62, 63)	0.34624261	2.42369747	2	01 23
3	NUMBER OF COMMERCIAL SUBJECTS	0.00000000	0.00000000	4	0 1 2 34
4	NUMBER OF NON-ACADEMIC SUBJECTS	0.04140285	0.04140285	2	0 1
5	TYPE OF SCHOOL	0.00000000	0.00000000	2	B C
6	MATHEMATICS MARK	0.00900733	0.09908056	2	FE ?DCBA
7	ENGLISH MARK	0.00000194	0.00002138	2	FE DCB?A

CHARACTERISTICS OF THE BEST PREDICTOR

3	NUMBER OF COMMERCIAL SUBJECTS	0.00000000	0.00000000	4	0 1 2 34

PREDICTOR: NUMBER OF COMMERCIAL SUBJECTS * ROW PERCENTAGES *

	0	1	2	3,4	TOTAL
OTHER	4.9	10.3	13.2	4.1	6.6
JMB	29.7	6.5	0.0	0.0	20.5
TSSC & TUEC	63.7	73.5	15.8	0.0	58.6
NSC	1.6	9.7	71.1	95.9	14.3
TOTALS (100%)	427	155	38	49	669

TABLE IS SIGNIFICANT AT 0.00000% LEVEL

(3)

SUMMARY OF RESULTS
ANALYSIS OF SUBGROUP NUMBER OF COMMERCIAL SUBJECTS (0)

NO.	NAME	SIGNIF (%)	BONF-SIG (%)	NUMBER OF GROUPS.	
1	DID STUDENT EVENTUALLY GET THE CTA (YES,NO)	100.00000000	100.00000000	1	YN
2	YEAR OF ENTRY (60, 61, 62, 63)	100.00000000	100.00000000	1	0231
3	NUMBER OF COMMERCIAL SUBJECTS	100.00000000	100.00000000	0	
4	NUMBER OF NON-ACADEMIC SUBJECTS	100.00000000	100.00000000	1	01
5	TYPE OF SCHOOL	0.00000435	0.00000435	2	B C
6	MATHEMATICS MARK	0.02579850	0.28378344	2	?FE DCBA
7	ENGLISH MARK	0.18634704	2.07182693	2	FE DCB?A

CHARACTERISTICS OF THE BEST PREDICTOR

5	TYPE OF SCHOOL	0.00000435	0.00000435	2	B C

PREDICTOR: TYPE OF SCHOOL * ROW PERCENTAGES *

	B	C	TOTAL
OTHER	5.1	0.0	4.9
JMB	27.7	86.7	29.7
TSSC & TUEC	66.0	0.0	63.7
NSC	1.2	13.3	1.6
TOTALS (100%)	412	15	427

TABLE IS SIGNIFICANT AT 0.00000% LEVEL

(4)

SUMMARY OF RESULTS
ANALYSIS OF SUBGROUP NUMBER OF COMMERCIAL SUBJECTS (1)

NO.	NAME	SIGNIF (%)	BONF-SIG (%)	NUMBER OF GROUPS.	
1	DID STUDENT EVENTUALLY GET THE CTA (YES,NO)	100.00000000	100.00000000	1	YN
2	YEAR OF ENTRY (60, 61, 62, 63)	100.00000000	100.00000000	1	0123
3	NUMBER OF COMMERCIAL SUBJECTS	100.00000000	100.00000000	0	
4	NUMBER OF NON-ACADEMIC SUBJECTS	100.00000000	100.00000000	1	01
5	TYPE OF SCHOOL	0.00000000	0.00000000	2	B C
6	MATHEMATICS MARK	100.00000000	100.00000000	1	EDC?BA
7	ENGLISH MARK	0.05186389	0.36304716	2	E DC?B

CHARACTERISTICS OF THE BEST PREDICTOR

5	TYPE OF SCHOOL	0.00000000	0.00000000	2	B C

****

PREDICTOR: TYPE OF SCHOOL * ROW PERCENTAGES *

	B	C	TOTAL
OTHER	11.4	0.0	10.3
JMB	5.7	13.3	6.5
TSSC & TUEC	80.0	13.3	73.5
NSC	2.9	73.3	9.7
TOTALS (100%)	140	15	155

TABLE IS SIGNIFICANT AT 0.00000% LEVEL

---

(5)

SUMMARY OF RESULTS
ANALYSIS OF SUBGROUP NUMBER OF COMMERCIAL SUBJECTS (0) TYPE OF SCHOOL (B)

NO.	NAME	SIGNIF (%)	BONF-SIG (%)	NUMBER OF GROUPS.	
1	DID STUDENT EVENTUALLY GET THE CTA (YES,NO)	100.00000000	100.00000000	1	YN
2	YEAR OF ENTRY (60, 61, 62, 63)	100.00000000	100.00000000	0	
3	NUMBER OF COMMERCIAL SUBJECTS	100.00000000	100.00000000	1	01
4	NUMBER OF NON-ACADEMIC SUBJECTS	100.00000000	100.00000000	0	
5	TYPE OF SCHOOL	0.03399508	0.37394583	2	?FE DCBA
6	MATHEMATICS MARK	0.27476025	2.47284222	2	E DCB?A
7	ENGLISH MARK				

CHARACTERISTICS OF THE BEST PREDICTOR

6	MATHEMATICS MARK	0.03399508	0.37394583	2	?FE DCBA

****

PREDICTOR: MATHEMATICS MARK * ROW PERCENTAGES *

	?,F,E	D,C,B,A	TOTAL
OTHER	9.8	3.8	5.1
JMB	41.3	23.8	27.7
TSSC & TUEC	47.8	71.3	65.0
NSC	1.1	1.3	1.2
TOTALS (100%)	92	320	412

TABLE IS SIGNIFICANT AT 0.03400% LEVEL

(6)

SUMMARY OF RESULTS
ANALYSIS OF SUBGROUP NUMBER OF COMMERCIAL SUBJECTS (I) TYPE OF SCHOOL (B)

NO.	NAME	SIGNIF (%)	BONF-SIG (%)	NUMBER OF GROUPS.	
1	DID STUDENT EVENTUALLY GET THE CTA (YES,NO)	100.00000000	100.00000000	1	YN
2	YEAR OF ENTRY (60, 61, 62, 63)	100.00000000	100.00000000	1	0123
3	NUMBER OF COMMERCIAL SUBJECTS	100.00000000	100.00000000	0	
4	NUMBER OF NON-ACADEMIC SUBJECTS	100.00000000	100.00000000	1	01
5	TYPE OF SCHOOL	100.00000000	100.00000000	0	
6	MATHEMATICS MARK	100.00000000	100.00000000	1	EDC?BA
7	ENGLISH MARK	0.60599446	4.24196053	2	E DC?B

CHARACTERISTICS OF THE BEST PREDICTOR

7	ENGLISH MARK	0.60599446	4.24196053	2	E DC?B

THIS PREDICTOR IS NOT SIGNIFICANT.

---

(7)

SUMMARY OF RESULTS
ANALYSIS OF SUBGROUP NUMBER OF COMMERCIAL SUBJECTS (O) TYPE OF SCHOOL (B) MATHEMATICS MARK (D,C,B,A)

NO.	NAME	SIGNIF (%)	BONF-SIG (%)	NUMBER OF GROUPS.	
1	DID STUDENT EVENTUALLY GET THE CTA (YES,NO)	100.00000000	100.00000000	1	YN
2	YEAR OF ENTRY (60, 61, 62, 63)	100.00000000	100.00000000	1	0123
3	NUMBER OF COMMERCIAL SUBJECTS	100.00000000	100.00000000	0	
4	NUMBER OF NON-ACADEMIC SUBJECTS	100.00000000	100.00000000	1	01
5	TYPE OF SCHOOL	100.00000000	100.00000000	0	
6	MATHEMATICS MARK	100.00000000	100.00000000	1	DCBA
7	ENGLISH MARK	100.00000000	100.00000000	1	EDCBA

NO PREDICTION IS POSSIBLE.

(8)

```
G V KASS. 1974. V.A.
ANALYSIS OF TOTAL GROUP.
**********
MONOTONIC PREDICTOR: DID STUDENT EVENTUALLY GET THE CTA (YES,NO)
TABLE HAS CHI-SQUARE   12.979, WITH D.F.   3, AND SIGNIF   0.48708707%

  2 GROUPS:                Y     N
TEST STATS FOR GROUPING:     13.0

MIN STAT IS 12.9788,  TO MERGE (Y) AND (N).                    D.F.:   3,  SIGNIFICANCE   0.4871%

  2 GROUPS:                Y     N
TEST STATS FOR SPLITTING:

NO COMBINED GROUPS TO SPLIT

BEST IS  2 GROUPS, WITH CHI-SQUARE   12.979  D.F.:   3.      1 IS THE BONF MULT, AND THE SIGNIFICANCE IS    0.48708707%
**********
**********
FREE PREDICTOR:  YEAR OF ENTRY (60, 61, 62, 63)
TABLE HAS CHI-SQUARE   19.086, WITH D.F.   9, AND SIGNIF   2.44549656%

MERGE PHASE    4 GROUPS:        0     1     2
                          1   2.2
                          2   4.6   3.3
                          3  12.7  11.7   3.0

MIN STAT IS  2.1532,  TO MERGE GROUPS NUMBERED  1 AND  2.       D.F.:   3,  SIGNIFICANCE  75.8551%

MERGE PHASE    3 GROUPS:        0;1     2
                          2      4.8
                          3     16.7   3.0

MIN STAT IS  2.9788,  TO MERGE GROUPS NUMBERED  2 AND  3.       D.F.:   3,  SIGNIFICANCE  61.0585%

MERGE PHASE    2 GROUPS:        0;1
                        2,3     13.7

MIN STAT IS 13.7450,  TO MERGE GROUPS NUMBERED  1 AND  2.       D.F.:   3,  SIGNIFICANCE   0.3462%

SPLIT PHASE    2 GROUPS:     ( 0     1 ) ( 2     3 )

TEST STATS FOR SPLITTING THE GROUP:    0 1
01   2.2

TEST STATS FOR SPLITTING THE GROUP:    2 3
01   3.0

MAX STAT IS  2.9788 TO SPLIT GROUP NUMBERED  2 AS 01.           D.F.:   3,  SIGNIFICANCE  61.0586%

BEST IS  2 GROUPS, WITH CHI-SQUARE   13.745  D.F.:   3.      7 IS THE BONF MULT, AND THE SIGNIFICANCE IS    0.34624261%
**********
```

(9)

CONTINUED

```
G V KASS. 1974. V.A.
ANALYSIS OF TOTAL GROUP.
**********
FREE PREDICTOR:  NUMBER OF COMMERCIAL SUBJECTS
TABLE HAS CHI-SQUARE  470.140, WITH D.F.  12, AND SIGNIF   0.00000000%

MERGE PHASE    5 GROUPS:        0     1     2     3
                          1   52.5
                          2 258.9  71.9
                          3 373.5 121.4   9.6
                          4 213.4  46.6   2.7   0.3

MIN STAT IS  0.3475,  TO MERGE GROUPS NUMBERED  4 AND  5.       D.F.:   3,  SIGNIFICANCE  99.9982%

MERGE PHASE    4 GROUPS:        0     1     2
                          1   52.5
                          2 258.9  71.9
                        3,4 390.2 132.0  11.5

MIN STAT IS 11.4039,  TO MERGE GROUPS NUMBERED  3 AND  4.       D.F.:   3,  SIGNIFICANCE   0.9499%

SPLIT PHASE    4 GROUPS:     ( 0 ) ( 1 ) ( 2 ) ( 3     4 )

TEST STATS FOR SPLITTING THE GROUP:    3 4
01   0.3

MAX STAT IS  0.3475 TO SPLIT GROUP NUMBERED  4 AS 01.           D.F.:   3,  SIGNIFICANCE  99.9982%

BEST IS  4 GROUPS, WITH CHI-SQUARE  469.839  D.F.:   9.      10 IS THE BONF MULT, AND THE SIGNIFICANCE IS    0.00000000%
**********
**********
FREE PREDICTOR:  NUMBER OF NON-ACADEMIC SUBJECTS
TABLE HAS CHI-SQUARE   18.562, WITH D.F.   3, AND SIGNIF   0.04140306%

MERGE PHASE    2 GROUPS:        0
                          1   18.6

MIN STAT IS 18.5621,  TO MERGE GROUPS NUMBERED  1 AND  2.       D.F.:   3,  SIGNIFICANCE   0.0414%

SPLIT PHASE    2 GROUPS:     ( 0 ) ( 1 )

NO COMBINED GROUPS TO SPLIT

BEST IS  2 GROUPS, WITH CHI-SQUARE   18.562  D.F.:   3.      1 IS THE BONF MULT, AND THE SIGNIFICANCE IS    0.04140285%
**********
**********
FREE PREDICTOR:  TYPE OF SCHOOL
TABLE HAS CHI-SQUARE  414.148, WITH D.F.   3, AND SIGNIF   0.00000000%

MERGE PHASE    2 GROUPS:        B
                          C  414.1

MIN STAT IS414.1482,  TO MERGE GROUPS NUMBERED  1 AND  2.       D.F.:   3,  SIGNIFICANCE   0.0000%
```

G V KASS. 1974. V.A.
ANALYSIS OF TOTAL GROUP.

(10)

CONTINUED

SPLIT PHASE 2 GROUPS: (B) (C)

NO COMBINED GROUPS TO SPLIT

BEST IS 2 GROUPS, WITH CHI-SQUARE 414.148 D.F.: 3. 1 IS THE BONF MULT, AND THE SIGNIFICANCE IS 0.00000000%

****

SINGLE FLOATING PREDICTOR: MATHEMATICS MARK
TABLE HAS CHI-SQUARE 42.130, WITH D.F. 18, AND SIGNIF 0.11044937%

7 GROUPS:	F	E	D	C	B	A
TEST STATS FOR GROUPING:		2.6	12.6	4.8	6.0	3.8
AND FOR GROUPING ? WITH:	1.9	2.9	1.7	5.2	1.9	5.1

MIN STAT IS 1.7445, TO MERGE (?) AND (D). D.F.: 3, SIGNIFICANCE 84.1319%

6 GROUPS:	F	E	(? | D)	C	B	A
TEST STATS FOR GROUPING:		2.6	13.0		5.3	6.0	3.8

MIN STAT IS 2.6449, TO MERGE (F) AND (E). D.F.: 3, SIGNIFICANCE 66.6316%

5 GROUPS:	(F	E)	(?	D)	C	B	A
TEST STATS FOR GROUPING:			13.4		5.3	6.0	3.8

MIN STAT IS 3.7685, TO MERGE (B) AND (A). D.F.: 3, SIGNIFICANCE 28.6929%

4 GROUPS:	(F	E)	(?	D)	C	(B	A)
TEST STATS FOR GROUPING:			13.4		5.3	7.3	

MIN STAT IS 5.2885, TO MERGE (?D) AND (C). D.F.: 3, SIGNIFICANCE 15.0001%

3 GROUPS:	(F	E)	(?	D	C)	(B	A)
TEST STATS FOR GROUPING:			14.9		6.3		

MIN STAT IS 6.3480, TO MERGE (?DC) AND (BA). D.F.: 3, SIGNIFICANCE 9.4256%

2 GROUPS:	(F	E)	(?	D	C	B	A)
TEST STATS FOR GROUPING:			22.1				

MIN STAT IS 22.0835, TO MERGE (FE) AND (?DCBA). D.F.: 3, SIGNIFICANCE 0.0090%

2 GROUPS:	(F	E)	(?	D	C	B	A	?)
TEST STATS FOR SPLITTING:	2.6		2.7	4.9	6.3	6.3	OMIT	
			OMIT	4.0	6.6	5.0		

MAX STAT IS 6.5826 TO SPLIT GROUP (?DCBA). D.F.: 3, SIGNIFICANCE 8.4970%

BEST IS 2 GROUPS, WITH CHI-SQUARE 22.083 D.F.: 3. 11 IS THE BONF MULT, AND THE SIGNIFICANCE IS 0.00900733%

****

(11)

CONTINUED

G V KASS. 1974. V.A.
ANALYSIS OF TOTAL GROUP.

****

SINGLE FLOATING PREDICTOR: ENGLISH MARK
TABLE HAS CHI-SQUARE 57.394, WITH D.F. 18, AND SIGNIF 0.00065686%

7 GROUPS:	F	E	D	C	B	A	
TEST STATS FOR GROUPING:	2.5	33.0	5.1	0.8	4.6		
AND FOR GROUPING ? WITH:	1.9	4.0	2.2	3.4	3.1	1.3	

MIN STAT IS 0.7579, TO MERGE (C) AND (B). D.F.: 3, SIGNIFICANCE 99.2059%

6 GROUPS:	F	E	D	(C	B)	A	
TEST STATS FOR GROUPING:	2.5	33.0	5.2		6.0		
AND FOR GROUPING ? WITH:	1.9	4.0	2.2		3.6	1.3	

MIN STAT IS 1.3333, TO MERGE (?) AND (A). D.F.: 3, SIGNIFICANCE 92.1104%

5 GROUPS:	F	E	D	(C	B)	(?	A)
TEST STATS FOR GROUPING:	2.5	33.0	5.2		7.1		

MIN STAT IS 2.4725, TO MERGE (F) AND (E). D.F.: 3, SIGNIFICANCE 69.7351%

4 GROUPS:	(F	E)	D	(C	B)	(?	A)
TEST STATS FOR GROUPING:		33.6	5.2		7.1		

MIN STAT IS 5.2030, TO MERGE (D) AND (CB). D.F.: 3, SIGNIFICANCE 15.5675%

3 GROUPS:	(F	E)	(D	C	B)	(?	A)
TEST STATS FOR GROUPING:		43.0			4.8		

MIN STAT IS 4.8299, TO MERGE (DCB) AND (?A). D.F.: 3, SIGNIFICANCE 18.2929%

2 GROUPS:	(F	E)	(D	C	B	?	A)
TEST STATS FOR GROUPING:		42.6					

MIN STAT IS 42.6080, TO MERGE (FE) AND (DCB?A). D.F.: 3, SIGNIFICANCE 0.0000%

2 GROUPS:	(F	E)	(?	D	C	B	A	?)	
TEST STATS FOR SPLITTING:	2.5		2.4	4.2	0.1	3.8	OMIT		
			OMIT	4.1	0.2	4.8			

MAX STAT IS 4.8299 TO SPLIT GROUP (DCB?A). D.F.: 3, SIGNIFICANCE 18.2929%

BEST IS 2 GROUPS, WITH CHI-SQUARE 42.608 D.F.: 3. 11 IS THE BONF MULT, AND THE SIGNIFICANCE IS 0.00009194%

****

BEST TO USE VAR 3 NUMBER OF COMMERCIAL SUBJECTS TO GIVE 4 NEW GROUPS

(12)

G V KASS. 1974. V.A.
ANALYSIS OF TOTAL GROUP. << BEFORE GROUPING >>

PREDICTOR: DID STUDENT EVENTUALLY GET THE CTA (YES,NO) * ROW PERCENTAGES *

	Y	N	TOTAL
OTHER	6.0	7.0	6.6
JMB	19.1	21.6	20.5
TSSC & TUEC	65.2	53.2	58.6
NSC	9.7	18.1	14.3
TOTALS (100%)	299	370	669

TABLE IS SIGNIFICANT AT 0.48709% LEVEL

PREDICTOR: YEAR OF ENTRY (60, 61, 62, 63) * ROW PERCENTAGES *

	0	1	2	3	TOTAL
OTHER	4.1	7.4	6.9	8.0	6.6
JMB	22.9	19.3	18.8	20.9	20.5
TSSC & TUEC	62.9	63.6	58.1	49.1	58.6
NSC	10.0	9.7	16.3	22.1	14.3
TOTALS (100%)	170	176	160	163	669

TABLE IS SIGNIFICANT AT 2.44550% LEVEL

PREDICTOR: NUMBER OF COMMERCIAL SUBJECTS * ROW PERCENTAGES *

	0	1	2	3	4	TOTAL
OTHER	4.9	10.3	13.2	4.8	0.0	6.6
JMB	29.7	6.5	0.0	0.0	0.0	20.5
TSSC & TUEC	63.7	73.5	15.8	0.0	0.0	58.6
NSC	1.6	9.7	71.1	95.2	100.0	14.3
TOTALS (100%)	427	155	38	42	7	669

TABLE IS SIGNIFICANT AT 0.00000% LEVEL

PREDICTOR: NUMBER OF NON-ACADEMIC SUBJECTS * ROW PERCENTAGES *

	0	1	TOTAL
OTHER	6.7	5.7	6.6
JMB	21.4	9.4	20.5
TSSC & TUEC	56.3	84.9	58.6
NSC	15.6	0.0	14.3
TOTALS (100%)	616	53	669

TABLE IS SIGNIFICANT AT 0.04140% LEVEL

PREDICTOR: TYPE OF SCHOOL * ROW PERCENTAGES *

	A	B	C	TOTAL
OTHER	0.0	7.2	3.0	6.6
JMB	0.0	21.4	15.2	20.5
TSSC & TUEC	0.0	68.4	2.0	58.6
NSC	0.0	3.0	79.8	14.3
TOTALS (100%)	0	570	99	669

TABLE IS SIGNIFICANT AT 0.00000% LEVEL

(13)

G V KASS. 1974. V.A.
ANALYSIS OF TOTAL GROUP. << BEFORE GROUPING >>

CONTINUED

PREDICTOR: MATHEMATICS MARK * ROW PERCENTAGES *

	?	F	E	D	C	B	A	TOTAL
OTHER	25.0	0.0	6.5	8.9	3.6	7.6	2.9	6.6
JMB	25.0	100.0	27.1	19.2	18.5	19.6	8.6	20.5
TSSC & TUEC	50.0	0.0	43.9	59.3	62.5	66.3	82.9	58.6
NSC	0.0	0.0	22.6	12.6	15.5	6.5	5.7	14.3
TOTALS (100%)	4	1	155	214	168	92	35	669

TABLE IS SIGNIFICANT AT 0.11045% LEVEL

PREDICTOR: ENGLISH MARK * ROW PERCENTAGES *

	?	F	E	D	C	B	A	TOTAL
OTHER	25.0	0.0	4.9	8.3	4.9	4.2	25.0	6.6
JMB	25.0	100.0	28.6	14.9	19.5	12.5	25.0	20.5
TSSC & TUEC	50.0	0.0	43.7	65.1	69.1	75.0	25.0	58.6
NSC	0.0	0.0	22.8	11.8	6.5	8.3	25.0	14.3
TOTALS (100%)	4	1	224	289	123	24	4	669

TABLE IS SIGNIFICANT AT 0.00066% LEVEL

(14)

G V KASS. 1974. V.A.
ANALYSIS OF TOTAL GROUP.

PREDICTOR: DID STUDENT EVENTUALLY GET THE CTA (YES,NO) * ROW PERCENTAGES *

	Y	N	TOTAL
OTHER	6.0	7.0	6.6
JMB	19.1	21.6	20.5
TSSC & TUEC	65.2	53.2	58.6
NSC	9.7	18.1	14.3
TOTALS (100%)	299	370	669

TABLE IS SIGNIFICANT AT 0.48709% LEVEL

PREDICTOR: YEAR OF ENTRY (60, 61, 62, 63) * ROW PERCENTAGES *

	0,1	2,3	TOTAL
OTHER	5.8	7.4	6.6
JMB	21.1	19.8	20.5
TSSC & TUEC	63.3	53.6	58.6
NSC	9.8	19.2	14.3
TOTALS (100%)	346	323	669

TABLE IS SIGNIFICANT AT 0.34624% LEVEL

PREDICTOR: NUMBER OF COMMERCIAL SUBJECTS * ROW PERCENTAGES *

	0	1	2	3,4	TOTAL
OTHER	4.9	10.3	13.2	4.1	6.6
JMB	29.7	6.5	0.0	0.0	20.5
TSSC & TUEC	63.7	73.5	15.8	0.0	58.6
NSC	1.6	9.7	71.1	95.9	14.3
TOTALS (100%)	427	155	38	49	669

TABLE IS SIGNIFICANT AT 0.00000% LEVEL

PREDICTOR: NUMBER OF NON-ACADEMIC SUBJECTS * ROW PERCENTAGES *

	0	1	TOTAL
OTHER	6.7	5.7	6.6
JMB	21.4	9.4	20.5
TSSC & TUEC	56.3	84.9	58.6
NSC	15.6	0.0	14.3
TOTALS (100%)	616	53	669

TABLE IS SIGNIFICANT AT 0.04140% LEVEL

PREDICTOR: TYPE OF SCHOOL * ROW PERCENTAGES *

	B	C	TOTAL
OTHER	7.2	3.0	6.6
JMB	21.4	15.2	20.5
TSSC & TUEC	68.4	2.0	58.6
NSC	3.0	79.8	14.3
TOTALS (100%)	570	99	669

TABLE IS SIGNIFICANT AT 0.00000% LEVEL

(15)

G V KASS. 1974. V.A.
ANALYSIS OF TOTAL GROUP.

CONTINUED

PREDICTOR: MATHEMATICS MARK * ROW PERCENTAGES *

	F,E	?,D,C,B,A	TOTAL
OTHER	6.4	6.6	6.6
JMB	27.6	18.3	20.5
TSSC & TUEC	43.6	63.2	58.6
NSC	22.4	11.9	14.3
TOTALS (100%)	156	513	669

TABLE IS SIGNIFICANT AT 0.00901% LEVEL

PREDICTOR: ENGLISH MARK * ROW PERCENTAGES *

	F,E	D,C,B,?,A	TOTAL
OTHER	4.9	7.4	6.6
JMB	28.9	16.2	20.5
TSSC & TUEC	43.6	66.2	58.6
NSC	22.7	10.1	14.3
TOTALS (100%)	225	444	669

TABLE IS SIGNIFICANT AT 0.00000% LEVEL