5-1971

The Effectiveness of Categorical Variables in Discriminant Function Analysis

Preston Jay Waite
Utah State University

THE EFFECTIVENESS OF CATEGORICAL VARIABLES IN DISCRIMINANT FUNCTION ANALYSIS

by

Preston Jay Waite

A thesis submitted in partial fulfillment
of the requirements for the degree

of

MASTER OF SCIENCE

in

Applied Statistics

Approved:

UTAH STATE UNIVERSITY
Logan, Utah

1971

ACKNOWLEDGEMENTS

I am greatly indebted to Dr. Rex L. Hurst for his help and suggestions in the preparation of this thesis. As deadlines approached and tempers wore thin, I found Dr. Hurst always willing to give of his time to aid in the completion of this paper. My association with Dr. Hurst has been a rich learning experience, one I feel very fortunate to have had.

I would also like to thank the members of my committee, Dr. Ronald V. Canfield and Dr. James Shaver for their contributions to this paper as well as to my education.

In addition I wish to thank Dr. Donald V. Sisson who, more than anyone else was responsible for my choosing to pursue a course of study in statistics.

Finally I wish to express a sincere gratitude to my wife and family for their unselfish support, devoted encouragement and patience during the time the thesis was being prepared.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

The Effectiveness of Categorical Variables

in Discriminant Function Analysis

by

Preston Jay Waite, Master of Science

Utah State University, 1971

Major Professor: Dr. Rex L. Hurst
Department: Applied Statistics

A preliminary study of the feasibility of using categorical variables in discriminant function analysis was preformed. Data including both continuous and categorical variables were used and predictive results examined.

The discriminant function techniques were found to be robust enough to include the use of categorical variables.

Some problems were encountered with using the trace criterion for selecting the most discriminating variables when these variables are categorical. No monotonic relationship was found to exist between the trace and the number of correct predictions.

This study did show that the use of categorical variables does have much potential as a statistical tool in classification procedures.

(50 pages)

CHAPTER I

INTRODUCTION

The technique of discriminant function analysis was originated by R.A. Fisher and was first applied by Barnard (1935). Fisher's first paper on the subject appeared in 1936 (Fisher, 1936). In this study, Fisher was using measurements made on the Iris plant to predict the various varieties of Iris. This work, by Fisher, was based on the multivariate normal, with all the variables continuous. From Fisher's work it is seen that the discriminant function analysis is basically a procedure for finding a linear function of variables which will predict the group membership of observations. Much of the data currently being collected are at least in part categorical in nature. In the area of multiple regression, the usage of categorical variables has been added to the multiple regression techniques by means of dummy variables such as indicated by Harvey (1964), Henderson (1953) and Searl (1966). The multiple regression techniques have been found robust enough so that the usage of categorical variables has become prevalent, and it was decided to see if the discriminant function techniques are robust enough to use dummy variables in predicting categories or group membership.

This study will concentrate on basically two questions:

1. Can categorical variables be used in discriminant function analysis?

2. What problems, if any, are encountered in interpretation of the results when categorical variables are used?

The first part of the study will give the mathematical derivation of discriminant function analysis and classification procedures.

This chapter will also indicate how the categorical variables have been added to the discriminant function procedure by the use of dummy variables.

Chapter III contains a discussion of the first question concerning the feasibility of using categorical variables in discriminant function analysis. To examine this problem, a sample problem is analyzed using data collected by the author from a dental health survey conducted in the State of Utah.

Chapter IV contains a detailed discussion of the behavior of dummy variables in discriminant function analysis.

Chapter V, the final chapter, is a summary of the study. In this chapter, areas of further research are suggested for consideration of the reader.

If categorical variables can be used in discriminant function analysis, then statistical techniques are available which will handle the building of mathematical models when the dependent variables are categorical and the independent variables are any mixture of qualitative and quanative variables, via discriminant function analysis.

CHAPTER 2

MATHEMATICAL FORMULATION

2.1 Discriminant function

The technique of discriminant function analysis was originated by R.A. Fisher and was first applied by Barnard (1935). Fisher's first paper on the subject appeared in 1936 (Fisher, 1936). Fisher (1936) defines discriminant function between two populations as that linear function of characters for which the ratio

$$(\text{mean difference})^2 \div \text{variance}$$

is a maximum.

Rao (1965) gives a more recent discussion of discriminant function analysis, which follows very closely the development of Fisher. The mathematical formulation in this study follows the procedures suggested by Rao.

Let $a_1x_{1i} + a_2x_{2i} + \ldots a_px_{pi}$ be the linear function and d_i the difference of the expected values of x_i in the two populations.

$$\text{mean difference} = \Sigma a_i\, E(x_{i1}) - a_i\, E(x_{i2}) = \sum_i a_i\,(E(x_{i1}) - E(x_{i2})) = \sum_i a_i d_k \qquad (2.1.1)$$

$$\text{variance} = \Sigma\Sigma a_i a_j \sigma_{ij} \qquad (2.1.2)$$

The quantity to be maximized is then

$$(\Sigma a_i d_i)^2 \div \Sigma\Sigma a_i a_j \sigma_{ij} \qquad (2.1.3)$$

differentiating with respect to a_i and setting the derivative equal to zero gives the result

$$a_i\sigma_{ij} + a_2\sigma_{i2} + \ldots a_p\sigma_{ip} = c\delta_i \qquad i = 1, \ldots p \qquad (2.1.4)$$

Since only the ratio of a_i can be uniquely determined, c can be chosen to be equal to 1.

The coefficients a_i can be estimated by solving the equations

$$\begin{aligned} a_1 s_{11} + a_2 s_{12} + \ldots a_p s_{1p} &= d_1 \\ a_2 s_{21} + a_2 s_{22} + \ldots a_p s_{2p} &= d_2 \end{aligned} \qquad (2.1.5)$$

where s_{ij} is the variance-covariance matrix for i and j.

Discriminant function is a useful tool of statistics in that it allows the researcher to reduce the dimensions of his problem without loosing a great deal of predictive power. The classification procedures to be discussed in this study do not depend on the use of discriminant function. Individuals could be classified directly from the observations. Using discriminant function makes interpretation of the classification procedure more straightforward. Procedures are also available (Miller, 1960) for obtaining a stepwise deletion of the variables.

2.2. Classification procedures using discriminant function

In using discriminant function for classification purposes, the problem becomes one of deciding on the membership of an individual in one of a given set of populations. An attempt is made to look at the entire profile of the individual and compare this with the profile of the various groups. In order to obtain a satisfactory solution to the problem, the following information is necessary.

1. Probability densities $P_1(z)$, $P_2(z)$, . . . $P_k(z)$ for a given set of measurements z on an individual in k alternative populations.

2. A priori probabilities $\pi_1, \pi_2, \pi_3, \ldots \pi_k$ for the populations.

3. Assignment of loss function, R_{ij}, for misclassification.

Given an individual with measurements z, his probability of being in population j, and having measurements z given he comes from population j, is a measure of interest. The assumption is made that $P(G_i)$, probability that an individual comes from population j, and $P(z_i/G_j)$, probability that an individual would have a measurement Z_i given that he came from population j, are independent. Therefore

$$P((G_j) \cap (Z_j/G_j)) = P(G_j) \cdot P(Z_j/G_j). \qquad (2.2.1)$$

Now multiplying each term by its appropriate loss function and taking the negative sum over all groups results in the equation

$$S_j = -[\pi_1 P_1(z) r_{1j} + \ldots \pi_k P_k(z) r_{kj}] \quad j = 1, \ldots k. \qquad (2.2.2)$$

S_j is then the individual's discriminant function score and the individual is classified in the group for which his score is the highest. Such a rule is shown by Rao (1965) to minimize the expected loss.

In many real world problems, the losses due to wrong classification may be difficult if not impossible to obtain. In such cases the best rule is to assign the individual with measurements Z to that population for which the posterior probability has the highest value. In this case the discriminant score for the j^{th} population is

$$S_j = \sum_{j=1}^{k} \pi_j P_j(z). \tag{2.2.5}$$

Consider the case where (Z) is distributed as a p-variate normal in each population. Choose $P_i(Z)$ as the normal density

$$(2\pi)^{-P/2} |\Sigma|_j^{-\frac{1}{2}} \exp\left(-\tfrac{1}{2}(Z-\mu_j)' \Sigma_j^{-1} (Z-\mu_j)\right) \pi_j = 1,2, \ldots k \tag{2.2.6}$$

with mean μ_j and dispersion matrix Σ_j.

Rao (1965) suggests taking the logarithm of $\pi_j P_j(Z)$ and omitting the factor $(2\pi)^{-p/2}$ common to all j and obtaining the equivalent discrimination score.

$$S_j = -\tfrac{1}{2} \log |\Sigma_j| - \tfrac{1}{2} (Z-\mu_j)' \Sigma_j^{-1} (Z-\mu_j) + \log \pi_j \tag{2.2.7}$$

consider

$$\tfrac{1}{2}(z-\mu_j)' \Sigma_j^{-1} (Z-\mu_j) \qquad (2.2.9)$$

which is distributed as χ^2 with p degrees of freedom.

Under the assumption that the populations do not differ in the dispersion matrices, equation 2.2.7 can be rewritten as

$$S_j = \tfrac{1}{2} \log |\Sigma| - \tfrac{1}{2}(Z-\mu_j)'\Sigma^{-1} (Z-\mu_j) + \log \pi_j \qquad (2.2.9)$$

where Σ is the pooled variance-covariance matrix.
2.2.8 can also be rewritten as

$$(Z-\mu_j)' \Sigma^{-1} (Z-\mu_j). \qquad (2.2.10)$$

This is still distributed as χ^2 with p degrees of freedom. Ommiting the parts of 2.2.9 not included in 2.2.10, will not alter the assignment of classification. Since

$$P(X/G_i) = P(Z-\mu)' \Sigma^{-1} (Z-\mu) \qquad (2.2.11)$$

evaluation of equation 2.2.10 will result in a χ^2 value for each group.

For classification analysis, we will assign the individual to the group for which $\chi^2 = (Z-\mu)' \Sigma^{-1} (Z-\mu)$ is a minimum, or equivalently, assign the individual to the group for which $P(X/G_i)$ is maximum.

In order to preform the classification, a discriminant function score is computed for each individual. These scores are compared with the profile of the corresponding scores of each group centroid. The individual is then assigned to that group with profile most resembling his own.

The discriminant function score for an individual for group j can then be calculated by

$$\sum_{i=1}^{n} a_{ij} z_{ij} \qquad j = 1,2, \ldots k \qquad (2.2.12)$$

where n is the number of variables measured and the a_{ij} may be obtained from the formulas

$$\begin{aligned} a_{ij} &= \sigma^{11} m_{ij} + \ldots \sigma^{1n} m_{nj} \\ a_{2j} &= \sigma^{21} m_{1j} + \ldots \sigma^{2n} m_{nj} \\ a_{nj} &= \sigma^{n1} m_{1j} + \ldots \sigma^{nn} m_{nj}. \end{aligned} \qquad (2.2.13)$$

The procedure recommended by Rao (1965) produces discriminant function scores which are highly correlated. Because of the high correlation of these discriminant function scores, it is difficult to interpret their meaning. Since this procedure produced scores which are highly correlated, it is called a nonorthogonal solution. An orthogonal solution to this problem has been proposed by researchers using discriminant function analysis in the social and behaviorial sciences. Cooley and Lohnes (1962) give a good discussion of this

approach. Orthogonal, as it is used here, refers to the fact that the discriminant function scores obtained by this method are uncorrelated. This procedure produces one less function than the number of groups or the number of variables whichever is smaller.

The discriminant functions in this problem are represented by the solution of

$$\underline{W}^{-1} \underline{B} \underline{V}_i = \lambda_i \underline{V}_i \qquad (2.2.14)$$

where the λ_i are the characteristic roots of $|\underline{W}^{-1} \underline{A} - \lambda_i I| = 0$ and $\underline{V}_i$ are the columns of the weighting coefficients matrix $\underline{V}$.

W = error variance-covariance matrix

A = group variance-covariance matrix

Group centroids $(\underline{G})$ are computed by the equation

$$\underline{G} = \underline{M}^{-1} \underline{V}. \qquad (2.2.15)$$

M is a vector of group means

V as defined above

The variance-covariance matrix of group centroids is given by

$$\Sigma = \underline{V}' W \underline{V}. \qquad (2.2.16)$$

All covariates should be zero since columns of $\underline{V}$ are orthogonal. These discriminant function scores form the basis for a decision concerning to which group an individual should be assigned.

The actual problem of classification reduces to one of testing a group of hypotheses regarding group membership. The likelihood of such a hypothesis may be written as $P(G_j/x_i)$ $i = 1,2, \ldots n$ and $j = 1,2, \ldots g$. There are g such hypotheses and the hypotheses with the highest probability is selected. Cooley and Lohnes (1962) refer to this as the maximum likelihood method of classification. An attempt is made to look at the entire profile of the individual and compare this with the profile of various groups. One way to describe such a distribution is in terms of ellipsis, each being the locus of points of a specified density. The size of the ellipse is determined by the value $z^2 = x_i' \underline{D}^{-1} x_i$ where D is the variance-covariance matrix and x_i is an m element vector of deviations.

$$x_i = [X_{1i} - \bar{X}_{1.}, X_{2i} - \bar{X}_{2.}, \ldots X_{ni} - \bar{X}_{n.}] \qquad (2.2.17)$$

As the values of z^2 increase, the density at that point $x_{12}, x_{22}, \ldots x_{n2}$ decreases. If points are selected at random from a multivariate normal, z^2 is distributed as χ^2 (chi-square) with p degrees of freedom. The χ^2 value is a measurement of the standardized distance in n space, that a point is from a given group centroid. One selection criterion then becomes: Classify the individual into the group with centroid whose normalized distance is nearest in the n dimensional sense to the individuals score, or select R_j such that

$$\chi_j^2 \leq \chi_k^2 \qquad k = 1,2, \ldots g.$$

If $(D_i = D_j)$ and the sizes of the g groups are equal, this decision rule will result in the minimum number of misclassifications (Cooley and Lohnes, 1962). The question entertained using this decision rule is: What is the probability that an observation from group j would lie this far from group j centroid?

Another way of looking at this problem is: Given an observation, what is the probability that it came from group j? This can be computed using Bayes's theorem which results in the equation.

$$P_{ij}\ (H_j/x_j) = \frac{\pi_j\ |D_j|^{-\frac{1}{2}}\ e^{-\frac{1}{2}\ x_j^2}}{\sum_k \pi_k\ |D_k|^{-\frac{1}{2}}\ e^{-\frac{1}{2}\ x_k^2}} \qquad (2.2.18)$$

$j = 1,2, \ldots g$

$i = 1,2, \ldots n.$

This second method of classification will generally give more accurate predictions when the a priori densities vary widely from uniform.

The classification procedures discussed earlier can be used for both the orthogonal and nonorthogonal procedures. The orthogonal procedure has the advantage of easier interpretation. For this reason, the orthogonal approach has been used on the data throughout this study. The approach has been to look at the discriminant function scores of an individual and compare the pattern of these individual scores with group centroids of the same scores for the sample data.

Some additional problems arise when the variables used to obtain these function scores are categorical in nature. One important underlying assumption in Fisher's derivation of the discriminant function techniques was that the independent variables were normally distributed. The generated dummy variables are by no means normally distributed. This problem is not as severe as one would originally fear. The Central Limit Theorem guarantees that if a fairly large number of variables are considered, the distribution of the discriminant function scores approaches the normal. The Central Limit Theorem becomes weaker and weaker as the number of variables decreases. It is also weakened when the preponderence of categorical variables in the final model increases. The situation will be most extreme when a model of only one categorical variable, with a small number of levels, is used.

2.3 Selection of variables

In discriminant function problems, a large number of independent variables are generally available. Frequently the number is too large to be considered practical. There are ways of reducing the number of independent variables while retaining maximum predictive power.

As a first step in reducing the number of variables to be considered, Hurst (1971) suggests computing a simple analysis of variance, completely randomized design, for continuous variables and a two-way independence χ^2 for categorical variable analysis. Variables that are not significant by these tests may be eliminated from further study. If a continuous variable does not show significance between group means, the frequency distributions will overlap to the extent that the variable will be useless for predictive purposes.

With the categorical variables, an attempt is made to develop relationships between these variables and another categorical variable, group membership. The two-way independence χ^2 tests for independence of pairs of categorical variables. Any categorical variable which is not significantly associated with group membership can not be expected to contribute discrimination information between groups.

This screening is not the limit of the available procedures. Many variables which show significance between groups may be giving the same information due to a high association between or among variables. Miller (1960) has shown that the trace of

$$Q = (\underline{W}^{-1} \underline{A}) \tag{2.3.1}$$

is a useful criterion for further reducing the number of variables to be used.

The trace of Q is computed as

$$\text{trace } (Q) = q_{11} + q_{22} + \cdots q_{kk}. \tag{2.3.2}$$

In the univariate case, this statistic reduces to the F ratio.

Hurst (1971) has modified the work of Shatzoff, Tsao and Fienberg (1968), wherein they give an alogrithm for computing all possible multiple regressions, to evaluate the trace of a matrix under stepwise conditions.

The trace is first evaluated with all variables in the model. Each variable is then successively removed from the model and the trace

is again evaluated with that variable removed*. The variable which causes the smallest decrease in the trace is then removed permanently and the entire process is repeated. This process continues until only one variable remains in the model thus allowing the researcher to discover, in order of importance, the most discriminating variables. The last variable to reamin in the model is the one which best discriminates between groups. This procedure is analogous to the stepwise deletion procedures used in multiple regression with R^2 as the selection criterion.

When the number of independent variables to be considered is not too large, the initial screening procedures can be omitted and all variables run on the stepwise procedure.

2.4 The usage of qualitative variables in discriminant function analysis

Categorical variables can be introduced into discriminant function analysis by means of dummy variables.

Consider the case where an experiment contains p continuous variables and q categorical variables. The discriminant function model can be written as

$$Y_{ilm} = \sum_{j=1}^{q} \alpha_{ij} x_{jlm} + \sum_{j=p+1}^{p+q} \sum_{k=1}^{nj} \alpha_{ijk} z_{jklm} \qquad (2.4.1)$$

where

x_{jlm} are the categorical variables

z_{jklm} are the dummy variables of non full rank

n_j is the number of levels of the j^{th} categorical variable

---

* The dummy variables associated with a categorical variable are treated as a subset for deletion and trace comparisons.

The dummy variables can be brought to full rank by imposing the condition

$$\sum_{k=1}^{n_j} \alpha_{ijk} = 0 \qquad (2.4.2)$$

For example: If x_1 is a qualitative variable with five possible levels or categories, the following dummy variable would result.

Level	Z_{11}	Z_{12}	Z_{13}	Z_{14}	Z_{15}
1	1	0	0	0	0
2	0	1	0	0	0
3	0	0	1	0	0
4	0	0	0	1	0
5	0	0	0	0	1

These Z variables produce a $Z'Z$ matrix of non full rank, but by imposing the conditions

$$\sum_{k=1}^{5} \alpha_{ijk} = 0$$

the following full rank variables are obtained.

X_{11}^*	X_{12}^*	X_{13}^*	X_{14}^*
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1
-1	-1	-1	-1

where $x_{jk}^* = Z_{jk} - Z_{jn_j}$ $\qquad k = 1,2, \ldots n_j-1.$

The model using dummy variables of full rank then becomes

$$Y_{ilm} = \sum_{j=1}^{p} \alpha_{ij} x_{jlm} + \sum_{j=p+1}^{p+q} \sum_{k=1}^{n_j-1} \alpha_{ijk}^* x_{jklm}^* \qquad (2.4.4)$$

and $$\alpha_{ijn_j} = - \sum_{k=1}^{n_j-1} \alpha_{ijk}^*. \qquad (2.4.5)$$

The total number of resulting variables is then

$$N = (P + \sum_{j=p+1}^{p+q} (n_j-1)). \qquad (2.4.6)$$

Once these dummy variables have been created they are used along with the continuous variables. The (n_j-1) dummy variables formed from a qualitative variable are treated as a subset representing the categorical variable.

CHAPTER 3

APPLICATION OF DISCRIMINANT ANALYSIS TO CATEGORICAL VARIABLES

3.1 Summary of the sample problem

The methodology for adding categorical variables to the discriminant function model has now been developed. Up to this point the use of categorical variables in discriminant function seems at least feasible. In order to test this further, some data were needed. Methodology which should work, but does not give accurate prediction, is of little use to the researcher.

The data for this example come from a dental survey of the State of Utah. This survey was sponsored by the Utah Foundation for Dental Health Education and Research. It was financed by the Utah State Division of Health. The data were collected by a team of pollsters under the direction of Rex L. Hurst. Since most of this data is of a categorical nature, it offers a good base for the examination of the use of categorical variables in discriminant function analysis.

The data used consist of eighteen variables from groups of sizes 56, 465, and 711. The three groups in this example represent individuals answering no response, yes, and no to the question: Would you favor a prepaid dental insurance plan in the State of Utah? Fourteen of the eighteen variables are categorical in nature and the remaining four are continuous.

Table 1 lists the eighteen variables as well as indicating which ones were selected for use in the final model. Eight of the nine variables selected were categorical in nature. Table 2 gives the codes and their meanings for these eight variables.

Table 1. Variables involved in original sample

	Variables	Codes	
1.	Oral health score	continuous	
2.	Sex	0-2	
3.	Age	continuous	*
4.	Number of missing teeth	continuous	
5.	Number of fillings	0-4	*
6.	Number of cavities	0-4	*
7.	Teeth replaced	0-2	
8.	Do you wear dentures	0-2	
9.	Visit dentist adults	0-6	*
10.	Visit dentist children	0-6 (no code 1)	*
11.	Why don't you visit dentist	0-4	
12.	Gum trouble	0-2	
13.	Occupation	0-9	*
14.	Hospital plan	0-2	*
15.	Cost of dental work	0-4	*
16.	Frequency of brushing	0-5	*
17.	Education of husband	0-5	
18.	Dental knowledge score	continuous	

*Indicates variables used in final model.

3.2 Analysis and interpretation of the sample problem

The data were analyzed using a group of programs available at Utah State University and called The Discriminant Function Package (Hurst 1971). This package consists of seven separate programs which, when used in combination, will provide an analysis of the type desired. The particular segments of the package used in this study were:

1. MACRDT
2. SDF
3. ODF
4. DFS

A stepwise deletion procedure was applied to the eighteen variables. Because of storage restrictions of the computer used to analyze the problem, the eighteen variables were arbitrarily divided into two groups. Each group was separately analyzed by the stepwise deletion criterion. The variables selected to go into each group can be found in Table 3 parts A and B. Following the stepwise procedure, the most discriminating variables from each group were then combined into a single group and a stepwise procedure was preformed on them again. These final nine variables were subsequently used in the model for prediction of group membership.

The results of this procedure are listed in Table 3 part c. Age, a continuous variable, was the last variable to be deleted. This indicates that if a researcher were allowed only one unit of information on an individual and from this was required to classify him, age would be the best unit to obtain. The other variables are also listed in order of their contribution to the trace of $(Q=\underline{W}^{-1}\underline{A})$.

Table 2. Codes and meanings for categorical variables

Occupation		
0= Non response	1= Manual labor	2= Clerical service
3= Trades	4= Farming	5= Housewife
6= Government	7= Owner-operator	8= Professional
Visit dentist children		
0= Non response	2= When have trouble	3= 2-3 years
4= Every year	5= Every six months	6= More often
Hospital plan		
0= Non response	1= Yes	2= No
Cost of last years dental		
0= Non response	1= $0-99	2= $100-199
3= $200-299	4= $300+	
Number of cavities		
0= Non response	1= None	2= 1-3
3= 4-6	4= 7+	
Visit dentist adults		
0= Non response	1= Never	2= When have trouble
3= 2-3 years	4= Every year	5= Every six months
Frequency of brushing		
0= Non response	1= Never	2= Seldom
3= Once a day	4= Twice a day	5= Three times or more a day
Number of fillings		
0= Non response	1= None	2= 1-9
3= 10-19	4= 20+	

Table 3. Results of stepwise deletion for selection of discriminating variables

A. Screening run Part I

Variable	Trace	
Gum trouble	179.58	
Dentures	177.44	
Oral health score	175.36	
Number of missing teeth	172.36	
Sex	170.07	
Why don't you visit dentist	166.76	
Teeth replaced	160.96	
Number of fillings	154.51	*
Visit dentist adults	146.92	*
Number of cavities	137.30	*
Visit dentist children	123.04	*
Age	100.20	*

B. Screening run Part II

Variable	Trace	
Dental knowledge score	125.98	
Education of husband	124.49	
Frequency of brushing	117.58	*
Hospital plan	107.67	*
Cost of dental work	96.67	*
Occupation	78.10	*

*Indicates variable selected to be used for final stepwise.

Table 3. Continued

C. Final run	
Variable	Trace
Number of fillings	216.78
Frequency of brushing	209.88
Visit dentist adults	201.82
Number of cavities	193.07
Cost of dental work	181.20
Hospital plan	165.60
Visit dentist children	147.84
Occupation	129.32
Age	100.20

When all variables involved are continuous, Miller (1960) has shown that contributions to trace and predictive power have a monotonic relationship. This means that the higher the contribution to the trace of a particular variable, the more predictive power the variable contains. A major purpose of this study was to find out if the trace criterion can be used as a measure of predictive power when using categorical variables. In order to investigate this problem, the discriminant function scores corresponding to the categorical variables were computed and the prediction results examined.

When the eight categorical variables are written with their dummy variable representation, they are expanded to 39 variables. These 39, plus the continuous variable age, result in 40 variables for the predictive model. The transformation to orthogonal discriminant

function space results in minimum (g-1, r) = two discriminant functions (Cooley and Lohnes, 1962).

The coeficients for the discriminant functions for the sample problem are listed in Table 4 parts A and B. These coeficients are analogous to be b_i's in regression analysis. An individual's score is computed as a linear combination of these coeficients. For example, if the average scores of Group One were relatively high while the average scores for Group Two were relatively low, the variables with higher positive coefficients would likely indicate a person from Group One.

The group centroids represent the average of all the discriminant function scores for a particular group on a given function. The profile of scores for a particular group centroid represent a pattern for that group. It is this pattern that is compared with the pattern of an individual in the classification procedures. The centours of the centroids are a measure of the overlap of the data from two different groups. The group centroids and centours of the centroids are listed in Table 5.

Figure 1 shows a two dimensional plot of the three group centroids for this problem. Some interpretation of what each of the scores mean based on the relative position of the centroids can now be made. The first score gives maximum discrimination between the No and Yes groups. The Non Response group can be looked at as being a composite of three groups.

Table 4. Individual coeficients for the discriminant function scores

	A. Coeficients for discriminant function one										
Variable	0	1	2	3	4	5	6	7	8	9	Cont.
Age											.018
Occupation	.322	.273	.045	.063	.154	-.129	-.173	-.093	.170	-.097	
Visit dentist children	-.025	----	-.004	.051	.011	.033	-.067				
Hospital plan	-.025	.128	-.102								
Cost of dental work	-.270	-.257	.049	.117	.361						
Number of cavities	.193	-.069	.076	.221	.421						
Visit dentist adults	-.588	-.123	-.928	.090	-.160	1.843					
Frequency of brushing	.264	.136	-.184	-.328	.034	-.089					
Number of fillings	.023	.025	.080	-.064	-.064						

Table 4. Continued

Variable	B. Coeficients for discriminant function two										
	0	1	2	3	4	5	6	7	8	9	Cont.
Age											-.001
Occupation	.038	.058	-.016	.045	.072	.188	.177	.381	-.978	.039	
Visit dentist children	.218	----	.136	.149	.154	.067	-.724				
Hospital plan	-.033	.055	-.023								
Cost of dental work	-.177	-.042	-.059	-.151	.430						
Number of cavities	.136	-.023	-.026	.099	-.184						
Visit dentist adults	.058	-.068	-.066	.088	-.383	.212					
Frequency of brushing	-.410	.280	.122	.137	.055	-.084					
Number of fillings	.183	.292	-.065	.084	.090						

Table 5. Group centroids and centours of the centroids

A. Group centroid for discriminant function one

Group	Centroid
1.	.8345
2.	.4680
3.	.9108

B. Group centroid for discriminant function two

Group	Centroid
1.	.3550
2.	.5801
3.	.5896

C. Centours of the centroids

1.0000	.2892	.5997
.2892	1.0000	.5547
.5997	.5547	1.0000

Those giving no response may be from one of the following groups.

1. In favor but choose not to answer.
2. Opposed but choose not to answer.
3. Really have no opinion.

As might be expected, the centroid of the Non Response group lies between the Yes and No groups on score one. The second discrimination score seems to be a score which tends to discriminate between response categories and the Non Response group.

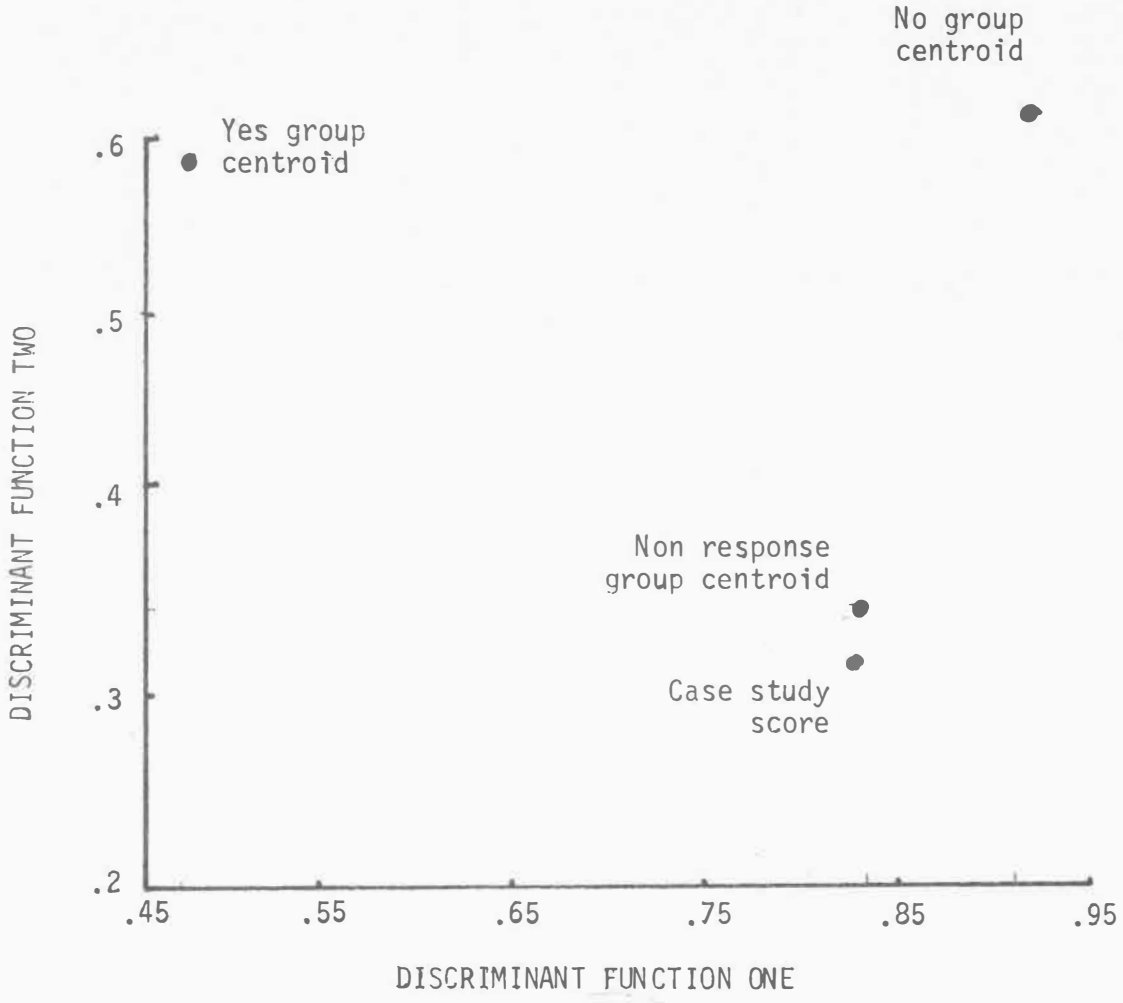

Figure 1. Relative position of group centroids on the two function scores.

By corresponding these scores to these meanings, some interpretation can be given to the coeficients in Table 4 parts A and B.

For discriminant function one, Groups One and Three represent individuals giving no response and those opposed to a dental insurance respectively. Both of these group centroids show a higher value. Those in favor show a somewhat lower centroid. The coefficients of Table 4 part A then can be given some interpretation. The higher postive coefficients for categories tend to classify the individual in either Non Response or No groups on discriminant function one. Consider, for example, the category, number of fillings. Table 2 indicates the five possible codes for this variable. In Table 4A, the coefficients for code three and four have negative coefficients. These coefficients indicate that based on discriminant function one alone, individuals who have more fillings would be more likely to favor the insurance.

The centroids in discriminant function two show that higher scores correspond with answering the questionaire, while lower scores correspond with non responses. In the question on hospital plan, those individuals who would not answer this question and those who had no hospital plan were generally non commitial on a dental plan.

An example of one respondent, and how he was classified, may be instructive to look at. This individual was actually from the group of Non Response, but individuals from any other group would be analyzed in the same manner. The data for this respondent are shown in Table 6.

Table 6. Data for individual case study

Variable number	Variable name	Response	Code	Dummy variable representation
1	Age	51	51	
2-10	Occupation	Trades	3	0 0 0 1 0 0 0 0 0 0
11-15	Visit dentist children	No response	0	1 0 0 0 0
16-17	Hospital plan	Yes	1	0 1
18-21	Cost of dental work	0-$99	1	0 1 0 0
22-25	Number of cavities	None	1	0 1 0 0
26-31	Visit dentist adult	2-3 Times/yr.	2	0 0 1 0 0 0
32-36	Frequency of brushing	3+	5	-1-1-1-1-1
37-40	Number of fillings	1-9	2	0 0 1 0

The two discriminant function scores for this individual are then computed by adding his variable score times the weighting factor for that category.

Score for discriminant function one.

$$51(.018) + 1(.063) + 1(-.025) + 1(.128) + 1(-.257)$$
$$+ 1(-.069) + 1(.090) + 1(-.089) + 1(.080) = .839$$

Score for discriminant function two.

$$51(.001) + 1(.045) + 1(.218) + 1(.055) + 1(.042)$$
$$+ 1(.023) + 1(.088) + 1(-.084) + 1(-.065) = .348$$

The three group centroids along with the respondents score are ploted in Figure 1. From looking at this figure it can readily be seen that this person would be placed in Group One by the minimum distance criterion discussed in Chapter II.

3.3 Classification results using nine variable model

To test the effectiveness of the categorical data in predicting of group membership, an attempt was made to classify the data into one of the three groups. This classification was preformed via the two classification criterion discussed in section 2.2.

1. A minimum χ^2 technique assigns the individual to the group for which $\chi^2_j = (Z-\mu_j)' \Sigma^{-1} (Z-\mu_j)$ is a minimum.

2. A Bayesian approach was also used which takes into account the a priori distribution of group membership. The a priori distributions used in this study were the sample sizes. If the sample sizes are not representative of the population probabilities, the population probabilities should be used.

Table 7 gives the results of the classification procedure. By use of the minimum χ^2, which assumes a uniform prior, 62 percent of the individuals were correctly predicted. Since the data varied substantially from uniform, the Bayesian approach would be expected to give better results. This is indeed the case. As Table 9 shows, 66 percent of the individuals were correctly predicted using this method. This represents an improvement of 100 percent over guessing, which would be expected to correctly predict 33 percent of the individuals. This adds credence to the use of categorical variables in discriminant function analysis.

Table 7. Compairson between actual and predicted responses on the nine variable model

A. Using minimum χ^2

		Predicted			
		NR	YES	NO	TOTAL
Actual	NR	25	15	16	56
	YES	49	333	83	465
	NO	65	243	403	711
	Total	139	691	502	1232

761 correct out of 1232 or 62 percent correct

B. Using Bayesian

		Predicted			
		NR	YES	NO	TOTAL
Actual	NR	9	15	32	56
	YES	11	275	179	465
	NO	18	159	534	711
	Total	38	449	745	1232

818 correct out of 1232 or 66 percent correct

The results of the two methods of selection for the case study discussed in section 3.2 are given in Table 8. This individual was placed in Group One using the minimum distance criterion. When the Bayesian appraoch was used the classification was switched to Group Three.

This shift in the second prediction is due to the fact that a small percentage of the data were actually from people from the first group. The low prior probability pushes the observation into one of the other groups.

Table 8. Probabilities and classification

	Group 1	Group 2	Group 3	Classification
Minimum χ^2 Probability	.9688	.6664	.8899	1
Bayesian	.0539	.3105	.6356	3

To give a more detailed picture of the behavior of the categorical variables, a new model was formed using only the three most discriminating variables. Two of these variables were categorical and one was continuous. As a final refinement of investigation, each of the three variables were used individually and in all possible combinations with the other two. Models were formed from each of these and a classification of every individual was made at each stage. Chapter IV gives a detailed breakdown of the results of these investigations.

CHAPTER IV

BEHAVIOR OF CATEGORICAL VARIABLES IN DISCRIMINANT ANALYSIS

4.1 Investigation procedures

In order to investigate the categorical data in detail, the three variables judged most discriminating by the trace criterion were selected for closer examination. One of these variables (age) is a continuous variable while the other two (occupation and visit dentist of children) are categorical. Table 9 gives the results of preliminary investigation of these three variables. The entries into the chi-square tables represent percentages of the group in each category and not actual numbers of observations.

As discussed in Chapter II, this univariate F test for the continuous variable age gives a measure of discrimination power for this variable. An F ratio of 100.2 is significant at the .001 level. A look at the means indicates why this is so high. Group three individuals, answering no, have the highest mean with a mean age of 49.01. This indicates that older people were more likely to be opposed to the plan. The mean of Group Two, those answering yes, is lower at 36.43. The Non Response group falls between the other groups as would be expected.

The two-way independence chi-square test was applied to the two categorical variables. This test gives some insight into where the discrimination power of the categorical variables occurs. Some levels, such as level one, indicate a wide disparity between groups. Based on occupation alone, those answering no response will likely come from group one.

Those in clerical-service occupations, column three of χ^2 table, are likely in Group Two. Other answers such as column one, manual labor, give little or no information on group membership.

Table 9. Results of univariate tests on the three most discriminating variables

Age			
Source of variation	Degrees of freedom	Mean squares	F test value
Total	1231		
Groups	2	22,585.3	100.2
Error	1229	225.4	

Group	Mean	Standard deviation
1.	47.66	17.90
2.	36.43	11.12
3.	49.01	16.86

Occupation										
	0	1	2	3	4	5	6	7	8	9
N = 56	.08	.00	.16	.24	.02	.00	.16	.05	.02	.27
N = 465	.02	.01	.25	.28	.02	.01	.19	.07	.07	.06
N = 711	.02	.02	.02	.23	.05	.04	.09	.04	.06	.29

$\chi^2_{18} = 147.379$

Table 9. Continued

	Visit dentist children					
	0	2	3	4	5	6
N = 56	.48	.02	.02	.17	.29	.02
N = 465	.30	.02	.01	.38	.28	.01
N = 711	.55	.02	.01	.18	.23	.01
				χ^2_{10} = 79.271		

A similar situation occurs with the third variable. Column 4 of the visit dentist children table, shows that those visiting the dentist every year, are likely to favor the insurance. Those not responding to this question are more likely to be opposed to the idea. Many individuals not responding to this question did so because they had no children living with them, which helps to explain why so many people in this category were opposed to the insurance.

As can be seen, all three variables have a large amount of discrimination information between groups. The trace criterion, shown in Table 3C, indicates that age is the most discriminating of the three variables.

If the three variables were to behave as continuous variables, then using age alone would result in more correct predictions than either of the other two variables alone. The results using categorical data did not seem to follow this pattern.

The procedure used to examine the behavior of categorical variables was to investigate the relative predictive power of the individual

variables alone as well as in combination with the other two. It was reasoned that by looking at a single categorical variable, and from that classifying all the individuals into one of the three groups, the effect of that variable alone could be examined. Since the trace criterion is available for these variables, a comparison of trace with ability to correctly predict could be made. In addition, investigation of combinations of variables would give an indication of how variables interact.

The fact that only the three most discriminating variables were thus examined is of no particular significance. Examination of more variables would have greatly magnified the analysis without providing a great deal of additional information.

4.2 Results and interpretation

All individuals in the study were classified using each model. A comparison of their power to correctly predict could then be made. Tables 10 and 11 give a complete breakdown of the successful predictions for each model.

These results tend to suggest that the trace criterion does not have a one to one correspondence with ability to correctly predict group membership. Using the minimum χ^2 criterion, age which should have been the best predicting single variable, correctly predicted only 407 individuals. Occupation, the variable with the second highest contribution to trace correctly predicted 679 individuals. This apparent contradiction can be explained, at least in part, without reference to the face that occupation is a categorical variable. The problem here is with the variable age. An extremely large group of individuals were incorrectly classified as coming from Group One.

This means that a large group of individuals, actually from Groups Two and Three were closer to the Group One centroid. From examining Figure 1, we see that the Group One centroid corresponds to lower values of discriminant function two than do the other two centroids. In Table 4B, coefficients for discriminant function two, the coefficient corresponding to age is .001. Using age alone, many individual scores would be low in function two, hence the misclassification.

Table 10. Comparison of the effectiveness of classification models using minimum χ^2

	Correctly predicted			
Actual group membership	Age alone	Occupation	Visit dentist children	Complete nine variable model
56	29	5	19	25
465	196	411	175	333
711	182	261	401	403
Total correct predictions	407	679	595	761

When the Bayesian approach is used, this problem is alleviated. Since the a priori probabilities of group one membership are so low, The Bayesian approach tends to force the individual into one of the other groups. When this is done the correct predictions increase to 759. This procedure is not a cure for all the problems. Using this criterion, the variable selected as third most discriminating correctly predicts more individuals than does the variable chosen second.

Table 11. Compairson of effectiveness of classification models using Bayesian procedure

Actual group membership	Correctly predicted							
	Age	Occupation	Visit dentist children	Full nine variable model	Age and occupation	Age and visit dentist children	Occupation and visit dentist children	Age and occupation and visit dentist children
56	0	0	0	9	0	0	1	0
465	208	248	175	275	247	270	210	276
711	551	496	581	534	538	504	564	528
Total correct predictions	759	744	756	818	785	783	775	804

Combinations of these three variables were then examined. The results did not agree with the results expected after seeing how the variables behaved by themselves.

The problem can best be analyzed using a two-dimensional example. The theory underlying the use of discriminant function requires that populations or groups in question follow a multivariate normal distribution. A one-dimensional picture may be helpful in illustrating the problem.

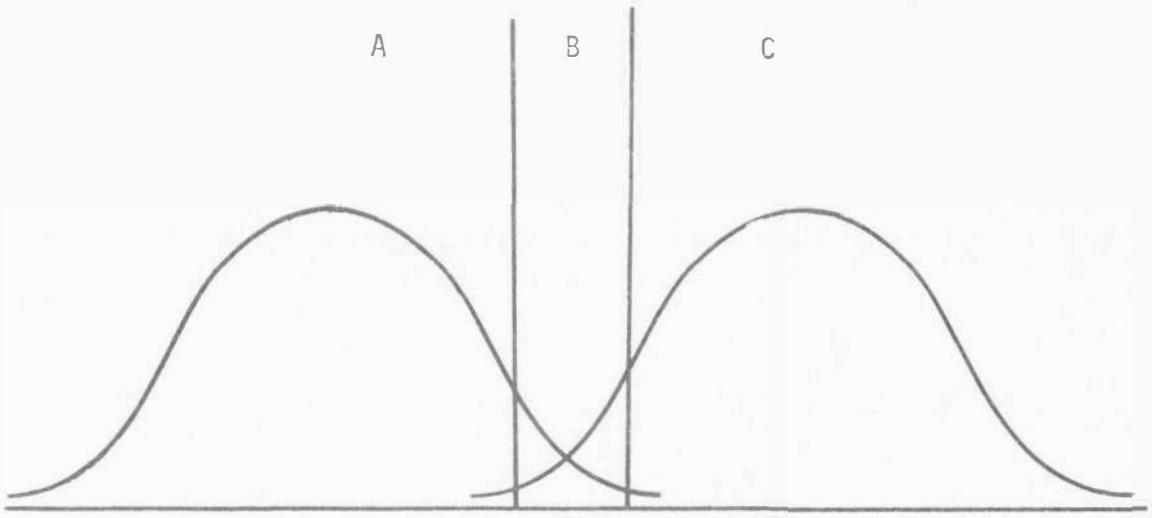

Figure 2. Regions of assigned group membership for continuous data.

This one dimension may be broken down into three areas. If an observation falls in region A, it will be classified as coming from Group I. If the observation falls in region C, it will be classified as coming from Group II. But what of those observations falling in region B? It is desired that this region of overlap be as small as possible. The trace criterion is in some sense a measure of this distance.

It is reasoned that the smaller this region for a particular variable the better is the power of predicting group membership.

When we are considering variables with good predictive power, the only portions of the curves of Groups I and II, lying in region B, are the tails of the multivarite normals. Therefore, if the selection criterion is moved slightly, this will cause a change in prediction of a very few observations. The trace has a monotonic relationship to the number of observations correctly predicted as long as the data are continuous.

When the case of categorical data is considered, the resulting distributions differ widely from the normal.

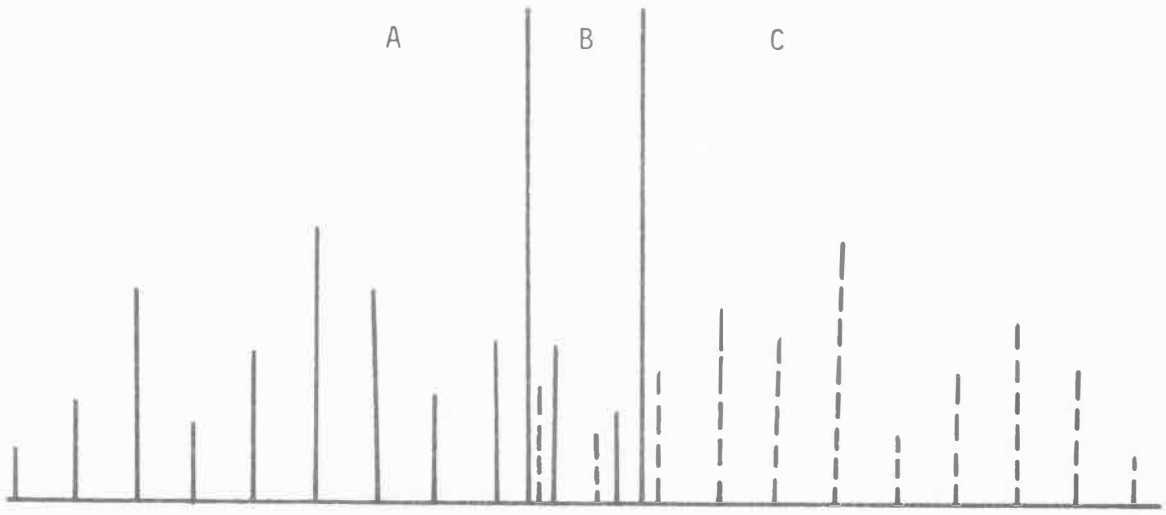

Figure 3. Regions of assigned group membership for categorical data.

The distributions are now discrete with several observations being clustered at only a few points. Now a very small change in selection criterion may cause the classification of several observations to be changed. The trace region still gives a measure of the width of B, but his width is no longer directly related to the power of correct prediction.

CHAPTER V

CONCLUSIONS AND FUTURE RESEARCH

The use of categorical variables in discriminant function analysis was shown to be an effective tool in the study of information. Many problems, however, still exist. The trace may not be the most effective tool in measuring discriminating information. The lack of a one-to-one correspondence makes omitting variables because of a small drop in trace, risky. At the present time, the author knows of no criterion which will give better results than does the trace. A rearcher desiring to use categorical variables should be aware of this problem when interpreting results. The only sure way of obtaining the unique best set of variables is to study all possible combinations and compare the results.

Future areas of study could include the derivation of a statistic similar to the trace which is not dependent on normality. This type of statistic should generalize to categorical data much easier. In addition some statistical test to indicate when a variable should remain in the model would be a real contribution. At present, the decision as to how many variables to keep for the final model is an arbitrary one with the researcher.

LITERATURE CITED

Baranad, M.M. 1935. The secular variations of skull characters in four series of Egyptian skulls, Annals of Egenics, 6:352-371.

Cooley, William W., and Lohnes, Paul R. 1962. Multivariate procedures for the behavioral sciences. John Wiley and Sons, Inc., New York, London.

Fisher, R.A. 1936. The use of multiple measurements in taxonomic problems, Annals of Egenics, 7:179-188.

Harvey, Walter R. 1964. Computing procedures for a generalized least-squares analysis program. U.S. Department of Agriculture, Agricultural Research Service, Biometrical Services.

Henderson, C.R. 1953. Estimation of general, specific and maternal combining abilities in crosses among inbred lines of swine. unpublished Ph.D. thesis. Iowa State College Library.

Hurst. Rex L. 1971. Discriminant function and classification analysis. unpublished paper, Department of Applied Statistics, Utah State University, Logan, Utah.

Miller, Robert G. 1960. Selecting variates for multiple discriminant analysis. FFCRC-TR-60-254, The Travelers Weather Research Center, Hartford, Connecticut.

Rao, C.R. 1965. Linear statistical inference and its applications. John Wiley and Sons, Inc., New York, London.

Searl, S.R. 1966. Matrix algebra for the biological sciences. John Wiley and Sons, New York, London.

Shatzoff, M., Tsao, R., and Fienberg, S. 1968. Efficient calculation of all possible regressions. Technometrics, 10:769-799.

VITA

Preston Jay Waite

Candidate for the Degree of

Master of Science

Thesis: The Effectiveness of Categorical Variables in Discriminant Function Analysis

Major Field: Applied Statistics

Biographical Information:

Personal Data: Born at Logan, Utah , April 2, 1946, son of W. Wayne and Faye S. Waite; married Judy Tolman June 10, 1968; one child--Stephanie.

Education: Attended elementary school in Hyde Park, Utah, graduated from North Cache High School in 1964; received the Bachelor of Science degree from Utah State University in Mathematics in 1970; completed requirements for the Master of Science degree, in Applied Statistics, at Utah State University in 1971.

Professional Experience: June 1970 to present, lecturer in Applied Statistics and Computer Science at Utah State University. June 1969 to present, Statistical Consultant for research workers on campus of Utah State University.