THE EGYPTIAN STATISTICAL JOURNAL
ISSR, Cairo Univ.,Vol.36,No.1,1992.

102

DISCRIMINANT ANALYSIS WITH MIXTURES OF CONTINUOUS, DISCRETE AND NOMINAL VARIABLES

A.M. KANDIL*

ABSTRACT

Discriminant analysis is studied in the case of mixed continuous, discrete and nominal variables. The probability of misclassification is derived and computations of numerical example is presented.

Key Words : Discriminant analysis, nominal variable, dummy binary variables, misclassifcation.

1. INTRODUCTION

This article considers the problem of discriminating between w-groups, and allocating individuals to one or another of these groups, when the available data consists of continuous, discrete and nominal variables.

The treatment of multivariate data has received substantial attention in the literature. If the variables are continuous, then they lead to the use of linear discriminant function, which was first, derived by Fisher (1936) and subsequently studied by many others (e.g. Anderson (1951, 1958); Welch (1939); Gilbert (1968, 1969); Lachenbruch, Sneeringer and Revo (1973) discussed the case when the variables are non-normal in particular, the case of binary data.

----

4. A NUMERICAL EXAMPLE

In this section, three training sets of data size n_1 = 500, corresponding to group H_1 (monthly income < 170), n_2 = 500, corresponding to group H_2 (monthly income 170-800), and n_3 = 500, corresponding to group H_3 (monthly income 800-2400), have beed taken as an example only to illustrate our procedure of reclassification outlined above. It should be noted that, these sets of data are derived from the data of industrialization and population project (Egyptian case study).

Suppose that, each data set consists of three different kinds of variables, one of which, Y as a response variable (amount income per month) is a continuous variable, and the others as explanatory variables, these explantory variables are:

(1) Z is a nominal variable corresponding to marital status, which takes 4 states [state (1) is married, (2) is unmarried, (3) is divorced, and (4) is widow (er)].

(2) X is a discrete variable corresponding to the number of family individuals (X = 0, 1, 2, ..., 8).

Now, the nominal variable can be replaced by 3-dummy binary variables. Let Z_1, Z_2, and Z_3 be these binary variables, each of which takes the value zero or one.

Then, state (1) is $Z_1 = 1$, $Z_2 = 0$, $Z_3 = 0$, state (2) is $Z_1 = 0$, $Z_2 = 1$, $Z_3 = 0$, state (3) is $Z_1 = 0$, $Z_2 = 0$, $Z_3 = 1$, and state (4) is $Z_1 = 0$, $Z_2 = 0$, $Z_3 = 0$.

The incidence table can be constructed with $\ell = 2^3 = 8$ cells from each training set, analogous to table (1). However, since it is impossible for Z_1 and Z_2 to be simultaneously equal to one and also, there is no multiple response, then one has to check that 4 cells in each table can not have any entries. This leaves 4 cells requiring estimated probabilities of occurence $\hat{P}_{lm}$ and probabilities of X and Y for each group.

The required computations for our model are:

(1) $P(H_1) = \hat{\pi}_1 = \frac{1}{3}$; $P(H_2) = \hat{\pi}_2 = \frac{1}{3}$; $P(H_3) = \hat{\pi}_3 = \frac{1}{3}$

(2) Estimated probabilities of occurrence $\hat{P}_{lm}$ are shown in table (2) . Note that, $\hat{P}_{lm} = \frac{n_{lm}}{n_l}$, $l = 1,2,3,\ m = 1,2,3,4,$

Table (2)

Estimated probabilities

of $\hat{P}_{lm}$

cell no. \ set no.	Set (1)	Set (2)	Set (3)
1	0.522	0.428	0.414
2	0.122	0.232	0.402
3	0.166	0.216	0.104
4	0.190	0.124	0.080
Sum	1.0	1.0	1.0

(3) Estimated parameters of X in each cell for each set $\hat{\phi}_{lm}$ are shown in table (3).

Table (3)

Estimated parameters of X

$\hat{\phi}_{lm}$

cell no. \ set no.	Set (1)	Set (2)	Set (3)
1	0.7509578	0.4918224	0.4088164
2	0.1250	0.125	0.125
3	0.4984939	0.386574	0.3629807
4	0.5078947	0.5846774	0.36250

(4) Estimated conditional probabilities $\hat{P}(X|Z)$ are shown in table (4).

Table (4)

Estimated probabilities $\hat{P}_i\ (X|Z)$ in each cell

for each training set

cell no.	value of X	Set (1)	Set (2)	Set (3)
(1)	0	0.000	0.004	0.015
	1	0.000	0.034	0.082
	2	0.004	0.117	0.200
	3	0.023	0.226	0.276
	4	0.086	0.273	0.239
	5	0.207	0.212	0.132
	6	0.311	0.103	0.046
	7	0.268	0.028	0.009
	8	0.101	0.003	0.001
(2)	0	0.344	0.344	0.344
	1	0.393	0.393	0.393
	2	0.196	0.196	0.196
	3	0.056	0.056	0.056
	4	0.010	0.010	0.010
	5	0.001	0.001	0.001
	6	0.000	0.000	0.000
	7	0.000	0.000	0.000
	8	0.000	0.000	0.000
(3)	0	0.004	0.019	0.027
	1	0.032	0.101	0.124
	2	0.111	0.223	0.247
	3	0.220	0.281	0.281
	4	0.273	0.222	0.200
	5	0.217	0.112	0.091
	6	0.108	0.035	0.026
	7	0.031	0.006	0.004
	8	0.004	0.001	0.000
(4)	0	0.003	0.001	0.027
	1	0.028	0.010	0.125
	2	0.103	0.049	0.247
	3	0.212	0.138	0.281
	4	0.273	0.243	0.199
	5	0.226	0.274	0.090
	6	0.117	0.193	0.026
	7	0.034	0.078	0.004
	8	0.004	0.014	0.001

(5) Estimated value of means for variable Y given value of X and value of Z for each training set and estimated value of common variance are shown in table (5)

Table (5)
Estimated value of $\hat{\mu}_{\cdot lmx_j}$ and $\hat{\sigma}_m^2$

cell no.	value of X	$\hat{\mu}_{1mx_j}$	$\hat{\mu}_{2mx_j}$	$\hat{\mu}_{3mx_j}$	$\hat{\sigma}_m^2$
(1)	0	0	0	0	$\hat{\sigma}_1^2$ =204863.25
	1	0	0	0	
	2	850	275.622	77.2	
	3	1161.278	414.921	102.956	
	4	1423.792	325.125	109.123	
	5	1487.821	311.827	105.182	
	6	1746.830	530.000	110.0	
	7	1461.50	225.769	94.0	
	8	1440.729	451.818	0.0	
(2)	0	0	0	0	$\hat{\sigma}_2^2$ =242624.12
	1	1507.377	430.75	84.055	
	2	0	0	0	
	3	0	0	0	
	4	0	0	0	
	5	0	0	0	
	6	0	0	0	
	7	0	0	0	
	8	0	0	0	
(3)	0	0	0	0	$\hat{\sigma}_3^2$ =243989.45
	1	2330.435	539.542	81.917	
	2	2156.500	527.875	84.357	
	3	2333.333	478.960	86.5	
	4	1682.000	270.000	119.85	
	5	1682.632	387.200	145	
	6	2611.111	470.000	107	
	7	2745.000	573.333	0	
	8	940.000	695.000	0	
(4)	0	0	0	0	$\hat{\sigma}_4^2$ =218842.99
	1	1243.040	540	109.923	
	2	2221.900	528.33	84.4	
	3	943.330	230	95	
	4	2441.539	643	155.5	
	5	2350.000	192.6	127.8	
	6	2237.143	460	100	
	7	2167.857	713.128	91	
	8	2433.33	754	0	

(6) Estimated valuees of $\hat{B}_{i\ mx_j}$, $\hat{C}_{i\ mx_j}$, and $\hat{\gamma}_{i\ mx_j}$ are shown in table (6)

Table (6)

Estimated values of

$\hat{B}_{i\nu mx_j}$, $\hat{C}_{i\nu mx_j}$ and $\hat{\gamma}_{i\nu mx_j}$

Cell no.	Value of x_j	$H_1 \dashrightarrow H_2$			$H_1 \dashrightarrow H_3$			$H_2 \dashrightarrow H_3$		
		$\hat{B}_{12mx_j}$	$-\hat{C}_{12mx_j}$	$\hat{\gamma}_{12mx_j}$	$\hat{B}_{13mx_j}$	$-\hat{C}_{23mx_j}$	$\hat{\gamma}_{13mx_j}$	$\hat{B}_{23mx_j}$	$-\hat{C}_{23mx_j}$	$\hat{\gamma}_{23mx_j}$
(1)	0	0	0	0	0	0	0	0	0	-1.2809
	1	0	0	0	0	0	0	0	0	-0.8339
	2	0.00280	1.5280	-3.1907	0.003772	1.7407	-3.7013	0.000969	0.1709	-0.5026
	3	0.00343	2.0712	-2.0936	0.005166	3.266	-2.2610	0.001523	0.3943	-0.1682
	4	0.00536	4.6097	-0.9547	0.006137	4.9106	-0.7905	0.001054	0.2290	0.16441
	5	0.00574	5.1653	0.1706	0.006749	5.3757	0.6806	0.001009	0.2164	0.50209
	6	0.005939	6.7619	1.2805	0.007961	7.4179	2.1411	0.002050	0.6561	0.0525
	7	0.006032	5.0888	2.4632	0.006675	5.1917	3.7160	0.00643	0.1029	1.2528
	8	0.004827	4.5678	3.9448	0.007033	5.661	5.0434	0.002206	0.4983	1.0986
(2)	0	0	0	-0.6433	0	0	-1.1942	0	0	-0.5509
	1	0.00437	0.3824	-0.6455	0.00437	0.0146	-1.1946	0.001429	0.3678	-0.5491
	2	0	0	-0.6533	0	0	-1.2024	0	0	-0.5491
	3	0	0	-0.6649	0	0	-1.2039	0	0	-0.5389
	4	0	0	-0.6931	0	0	-1.2039	0	0	-0.5108
	5	0	0	-0.6433	0	0	-1.1886	0	0	-0.5454
	6	0	0	0	0	0	0	0	0	0
	7	0	0	0	0	0	0	0	0	0
	8	0	0	0	0	0	0	0	0	0
(3)	0	0	0	-1.7047	0	0	-1.3063	0	0	0.3185
	1	0.00734	10.5329	-1.4214	0.00922	11.1157	-0.8873	0.00188	0.5828	-0.0487
	2	0.00667	0.9591	-0.6979	0.00849	9.5156	-0.3277	0.00182	0.5565	0.0838
	3	0.00760	10.6820	-0.5088	0.00921	11.1418	0.2283	0.00161	0.4548	0.2823
	4	0.00579	5.6483	-0.0558	0.00640	5.7682	0.7967	0.00615	0.11965	0.7329
	5	0.00531	5.4948	0.4008	0.00630	5.7509	1.3558	0.000993	0.2642	0.6913
	6	0.00878	13.5191	0.8755	0.0103	13.9483	1.9253	0.00149	0.4292	0.6206
	7	0.00890	14.7677	1.5404	0.011[illegible]	15.4413	2.6391	0.00235	0.6736	0.4250
	8	0.001004	0.8209	1.13079	0.00385	1.8108	7.4955	0.00285	0.9899	5.7998
(4)	0	0	0	1.5198	0	0	-3.6458	0	0	-2.8631
	1	0.00321	2.0641	1.3218	0.00518	3.5027	-0.6242	0.001965	0.6886	-2.7035
	2	0.00774	10.6417	1.1741	0.00977	11.2632	-0.0180	0.002029	0.6215	-1.8136
	3	0.00326	1.9123	0.8443	0.00388	2.0126	0.5931	0.000622	0.1002	-0.3515
	4	0.00822	12.7649	0.5361	0.00104	13.5644	1.1838	0.002228	0.8894	-0.2417
	5	0.00986	10.4765	0.2397	0.00102	12.5802	1.1083	0.000296	0.0474	1.5002
	6	0.00812	10.9513	-0.0764	0.00977	11.4119	2.3665	0.001645	0.4606	1.9809
	7	0.0066[illegible]	9.5723	-0.4418	0.00949	10.7185	3.0012	0.002853	1.1432	2.6253
	8	0.00767	12.2293	1.4035	0.01112	13.5282	4.5555	0.003445	1.2989	1.7723

(7) Estimated values of both the Mahalanobis squared distance $\hat{D}^2_{mx_j}$ and the cumulative standard normal distribution function $\hat{F}(u)$ are shown in table (7).

Table (7)

Estimated values of $\hat{D}^2_{mx_j}$ and $\hat{F}(u)$

cell n0.	value of x	D_{12mx_j}	D_{13mx_j}	D_{23mx_j}	$F_1(u)$	$F_2(u)$	$F_3(u)$
	0	0	0	0	0	0	0
	1	0	0	0	0	0	0
	2	1.61039	1.9152122	0.1921832	0.2743	0.1977	0.4129
	3	2.71913	5.467284	0.4750317	0.2061	0.1210	0.3632
(1)	4	5.89207	8.4366259	0.2277463	0.1131	0 07353	0.4052
	5	6.750314	9.3315448	0.2085027	0.0968	0.06301	0.4090
	6	7.22763	13.078053	0.8610622	0.09012	0.03515	0.3228
	7	7.453905	9.1283151	0.0847544	0.08534	0.06552	0.4404
	8	4.77365	10.132125	0.9964672	0.1379	0.04551	0.3085
	0	0	0	0	0	0	0
	1	4.77745	8.3497284	0.4954059	0.1379	0.07353	0.3632
	2	0	0	0	0	0	0
	3	0	0	0	0	0	0
(2)	4	0	0	0	0	0	0
	5	0	0	0	0	0	0
	6	0	0	0	0	0	0
	7	0	0	0	0	0	0
	8	0	0	0	0	0	0
	0	0	0	0	0	0	0
	1	13.14523	20.721524	0.8583184	0.03515	0.01130	0.3228
	2	10.871041	17.598206	0.806216	0.04947	0.01786	0.3264
	3	14.094551	20.690479	0.6312766	0.03005	0.01160	0.3446
(3)	4	8.171435	10.00133	0.0924016	0.07636	0.05705	0.4404
	5	6.87794	9.6902231	0.2404236	0.09510	0.05938	0.4014
	6	18.78916	25.700176	0.5400602	0.01500	0.005545	0.3557
	7	19.317397	30.882585	1.3472334	0.01390	0.002718	0.2810
	8	0.246015	3.6214681	1.9796963	0.4014	0.1711	0.2420
	0	0	0	0	0	0	0
	1	2.258538	5.8670104	0.8452006	0.2266	0.1131	0.3228
	2	13.106059	20.877554	0.9005383	0.03515	0.01130	0.3192
	3	2.325155	3.2885169	0.0845172	0.2236	0.1814	0.4404
(4)	4	14.78111	23.880017	1.0859669	0.02743	0.007344	0.3015
	5	21.26810	22.564912	0.0191874	0.01044	0.008656	0.4721
	6	14.43151	20.87058	0.5922054	0.02872	0.01130	0.3483
	7	9.669395	19.709724	1.7688958	0.05938	0.01321	0.2514
	8	12.8[illegible]6633	27.056361	2.597825	0.03673	0.004661	0.2090

Using the computations shown in tables (6) and (7) and applying models from (2.6) to (2.13) and the models (3.1) ,(3.2) , and (3.3) , we can find the estimate of η_{imx_j} and consequently $P(H_i|\zeta)$; $i = 1,2,$ and 3 ; $m = 1,2,$ 3 and 4 and $x_j = 0,1,\ldots 8$, for each indivaduals drawn from the three training sets , required for the allocation rule to classify these individuals ,we find that:

(1) 90 out of the 500 individuals , in group H_1 (18 percent) and 100 out of 500 individuals in group H_2 (20 percent) were misclassified . It means that 18 % of individuals drawn from H_1 actually belong to H_2 , and 20% of individuals drawn from H_2 actually belong to H_1 .

(2) 193 out of the 500 individuals in group H_2 (38.6 percent) and 147 out of 500 individuals in group H_3 (29.4 percent) were misclassified . It means that 38.6 % of individuals drawn from H_2 actually belong to H_3 and 29.4 % of individuals drawn from H_3 actually belong to H_2.

(3) 30 out of the 500 individuals in group H_1 (6 percent) were misclassified . It means that 6 % of individuals drawn H_1 actually belong to H_3.

(4) Assuming that the costs of misclassification are equal for the three training sets (note that it is not necessary at all) , then , by substituting into(3.2),(3.3), and (3.4) and using the computations shown in table (7) , we have the probabilities of misclassification as shown in table (8)

Table (8)

Estimated probabilities of misclassification on $\hat{P}(\nu|i)$ and $\hat{P}(i|\nu)$

cell n .	P (2|1)	P (1|2)	P (3|1)	P (1|3)	P(3|2)	P (2|3)
1	0.0526276	0.0608366	0.0294039	0.0418403	0.16015	0.145985
2	0.0066054	0.0125764	0.0035221	0.0116177	0.03332	0.057386
3	0.0096816	0.0111618	0.0057598	0.0033902	0.0776615	0.0367287
4	0.0138284	0.006831	0.0099989	0.0424151	0.0461503	0.0293472
sum	0.082745	0.0914058	0.0486947	0.0992333	0.3172818	0.2694462

It should be noted. that, the average of misclassification between H_1 and H_2 is 0.08707, the average of misclassification between H_1 and H_3 is 0.07396, and the average of misclassification between H_2 and H_3 is 0.29336.

(5) Note that the idea in this article can be extended and used to re-classify the egyptian income tax payers.

ACKNOWLEDGEMENT

I am grateful to Prof. D. Fareed El-Hossainee for his helpful comments. I am also, grateful to Dr. El-Saidi for revising the paper linguistically and I should like to thank the referee for his careful reading and useful comments.

REFERENCES

(1) Aitchison, J., and Aitken, C.,G.. G., (1976), "Multivariate Binary Discriminantion by the Kernal Method" Biometrika 63, 413-420.

(2) Anderson, J.A., (1972). "Separate Sample Logistic Discrimination". Biometeika, 59, 19-35.

(3) Anderson, J.A.. (1975) "Quadratic Logistic Discrimination" Bilometrika 62, 149-154.

(4) Anderson, T.,W., (1951) "Classification by Multivariate Analysis" Psychometrika 16, 31-50.

(5) Anderson, T., W., (1958) "An Introduction to Multivariate Statistical Analysis" New York: John Wiley & Sons, Inc.

(6) Change, P.C., and Afifi, A., A., (1974) "Classification Based on Dichotomous and Continuous Variables" Journal of American Statistical Association 69, 336-339.

(7) Day, N.E., and Kerridge, D.F., (1967) "A general Maximum likelihood Discriminat" Biometrics 23, 313-324.

(8) Fisher., R.A., (1936) "The Use of Multiple Measurements in Taxonomic Problems" Annals of Eugenics 7, 179-188.

(9) Gilbert, E.S., (1968) "On Discriminantion Using Qualitative Variables" J. of American Statistical Association 63, 1399-1412.

(10) Gilbert, E.S., (1969) "The Effect of Unequal Variance-Covariance Matrices on Fisher's linear Discriminant Function" Biometrics 25, 205-216.

(11) Hoel, P., and Petersen (1949) "A Solution to the Problem of Optimum Classification" Ann' Math. Statist. 20, 433-438.

(12) Krzanowski, W.J., (1975) "Discrimination and Classification Using Both Binary and Continuous Variables" J. of American Statistical Association 70, 782-790.

(13) Krzanowski, W.J., (1980) "Mixture of Continuous and Categorical Variables in Discriminant Analysis " Technometrics 10, 1-11.

(14) Krzanowski, W.,J., (1982) "Mixtures of Continuous and Categorial in Discriminant Analysis: A Hypothesis-Testing Approach" Biometrics 38, 991-1002.

(15) Lachenbruch, P., A., "Sneeringer, C. and Revo, L., T., (1937) "Robustness of the linear and Quadritic Discriminant Function to Certain Types of Nonnormality" Communications in Statistics, 1,39-56.

(16) Welch, B., L., (1939) "Note on Discriminant Functions" Biometrika 31, 218-220.

However, recent research has opened up the possibility of combining the different types of variables. Three distinct approaches to the treatment of mxied binary and continuous variables in discriminant analysis are evident in recent literature. Aitchison and Aitken (1976) suggested a method based on the kernal approach. Anderson (1972, 1975) discussed the use of logistic discrimination, in which the probability of group membership is assumed to be a logistic function of the observed variables.

Krzanowski (1975) studied the method of likelihood ratio based on the location model. He also discussed (Krzanowski (1980, 1982) the case of mixed continuous and categorical variables in discriminant analysis. The case of mixed continuous, discrete, and nominal varinbles in discriminant analysis does not seen to have received attention. The purpose of this paper is to study this case.

The model and an allocation rule are introduced in section (2). The probability of misclassification is discussed in section (3) and computations of numerical example based on hypothetical data, which consists of mixed variables (one continuous, one discrete, and one nominal variable) is presented in section (4).

(2) THE MODEL AND AN ALLOCATION RULE

Suppose that discrimination between W-groups H_1, H_2,..., H_w, is to be based on available W-sets of n_i observations known to have come from H_i groups (i=1,2,...,w). These sets are often referred to as training sets. We wish to set up a discrimination rule between H_i groups (i=1,2...,w) on the basis of three different kinds of vectors, observed on each observation. The first one, $Z = (Z_1 , Z_2, \ldots, Z_s)'$ is g-component vector of nominal variables, where the r th nominal variable has k_r states. The second vector, $X = (X_1, X_2, \ldots, X_t)$ is t-component vector of discrete variables, which has a multinomial distribution with parameters $(n_{ix_j}, \phi_{ix_j}$ $(i = 1,2,\ldots,w; j = 1,2,\ldots,t;$ and $X_j = 0,1,2,\ldots,n_j))$. The third one, $Y = (Y_1, Y_2, \ldots, Y_s)$ is S-Component vector of continuous variables, which has a multivariate normal distribution with parameters $(\mu_{ix_j}, \Sigma_{ix_j})$. Now we can replaced each nominal variable by (K_r-1)dummy binary variables, all these binary variables take the value zero, except the r th, which takes value one, if the corresponding nominal variable is observed in its r th state, $(r = 1,2,\ldots,k_r-1)$. Note that, all binary variables are zero for a nominal variable in k_r th state. Thus, if we have g-nominal variables, each of which has k_r states, then we replace it by $g(k_r - 1)$ dummy binary variables, each of which takes the value zero or one.

Thuse, these binary variables can be treated as a multinomial with $\ell = 2^{g(k_1-1)}$ states, and we can construct an incidence table with ℓ-cells from each training set. To illustrate this, suppose we have for example, two nominal variables, each with three states. Thus, each nominal variable is replaced by two dummy binary variables. So we have the total of 4 binary variables. Z_1 and Z_2 are two binary variables corresponding to the first one. Z_3 and Z_4 are the two binary variables corresponding to the second one. If the three states of the first variable are coded 1, 2, and 3, then state (1) becomes $Z_1 = 0$, $Z_2 = 1$; state (2) becomes $Z_1 = 1$, $Z_2 = 0$ and state (3) becomes $Z_1 = 0$, $Z_2 = 0$,. Similary for Z_3 and Z_4. The resulting incidence table for these cells is shown in table (1). It should be noted that 4-binary variables mean that there are $2^{g(k_1-1)} = 2^4 = 16$ cells in the multinomial table for each group.

TABLE (1)

AN INCIDENCE TABLE FOR 4 BINARY VARIABLES

cell no.	The first nominal variable				The second nominal variable			
	Z_1		Z_2		Z_3		Z_4	
	0	1	0	1	0	1	0	1
(1)	0	0	0	0	0	0	0	1
(2)	0	0	0	0	0	1	0	0
(3)	0	0	0	0	0	0	0	0
(4)	0	0	0	0	0	1	0	1
(5)	0	0	0	1	0	0	0	1
(6)	0	0	0	1	0	1	0	0
(7)	0	0	0	1	0	0	0	0
(8)	0	0		1	0	1	0	1
(9)	0	1	0	0	0	0	0	1
(10)	0	1	0	0	0	1	0	0
(11)	0	1	0	0	0	0	0	0
(12)	0	1	0	0	0	1	0	1
(13)	0	1	0	1	0	0	0	1
(14)	0	1	0	1	0	1	0	0
(15)	0	1	0	1	0	0	0	0
(16)	0	1	0	1	0	1	0	1

Let the probability of occurrence of cell m in group H_i be denoted by P_{im} (i = 1, 2, ..., w; m = 1, 2, ... ℓ) i. e. $P_i(Z) = P_{im}$: where $\sum_{m=1} P_{im} = 1$; and $0 \leq P_{im} \leq 1$ where i = 1, 2, ... w, m = 1, 2, ..., ℓ.

and $\ell = 2^{g(k_1-1)}$ ------ (2.1)

A suggested location model assumes that the conditional distribution of X in multinomial cell m is $P_i(X/Z)$, where,

$$P_1(X/Z) = \prod_{j=1}^{t} \frac{n_{1j}!}{x_{1mj}!} \phi_{1m_j}^{x_{1m_j}} \qquad \text{.............. (2.2)}$$

where $0 \leq \phi_{1mx_j} \leq 1$ is the parameter of x_j in cell m for group H_i and i = 1, 2, ..., w, j = 1, 2, ..., t, m = 1, 2, ... ℓ and x_j = 1, 2 ... n_j and the conditional distribution of Y given X and Z is a multivariate normal distribution with mean vetor μ_{1mx_j} and the common dispersion matrix Σ_{mx_j}

i.e., $f(Y/X,Z) \sim N(\mu_{1mx_j}, \Sigma_{mx_j})$ .... (2.3)

Now, let Π_1 denote the probability that the observation comes from the i <u>th</u> group, i.e.,

$P(H_i) = \Pi_1$ = 1, 2, ..., W and $\sum_{I=1}^{W} \Pi_1 = 1$ (2.4)

With this backgroud, we now construct the allocation rule which is based on Bayes's theorem (Hoel and Peterson 1949).

Since, the problem is to classify an boservation $\zeta = (Z, X, Y)$ into one of W- froups H_i (i = 1, 2, ..., W), and if ζ is from group H_i, then its density function (Chang and Afifi 1974) is,

$$f_i(\zeta) = f_i(z)\, f_i(x/z)\, f_i(Y/X,Z)$$

$$= \prod_{j=1}^{t} \alpha_i \frac{\grave{n}_{ij}\,!}{x_{imj}\,!} P_{im}\, \pi_i\, \phi_{imx_j}^{x_{imj}}\, \mathcal{C}(\mu_{imx_j}, \Sigma_{mx_j}) \qquad (2.5)$$

$i = 1,2, \ldots, W;\ j = 1,2, \ldots, t,\ m = 1,2, \ldots, \ell$ and

$x_j = 0,1,2, \ldots \grave{n}_j$

Where α_i is a constant chosen to make the total prbability unit.

Thus, the probability of the group H_i given ζ (Day and Kerridge 1967)

$$P(H_i/\zeta) = \frac{P(H_i)\, P(\zeta/H_i)}{\sum_{i=1}^{v} P(H_i)\, P(\zeta/H_i)} \qquad (2.6)$$

Appling the general model specified by (2.4) and (2.5) we find

$$P(H_i/\zeta) = \frac{\prod_{j=1}^{t} \alpha_i\, \beta_{ij}\, \pi_i\, P_{im}\, \phi_{imx_j}^{x_{imj}} \exp[-\tfrac{1}{2}(Y-\mu_{imx_j})' \Sigma_{mxj}^{-1} (Y-\mu_{imx_j})]}{\sum_{i=1}^{v} \prod_{j=1}^{t} \alpha_i\, \beta_{ij}\, \pi_i\, P_{im}\, \phi_{imx_j}^{x_{imj}} \exp[-\tfrac{1}{2}(Y-\mu_{imx_j})' \Sigma_{mxj}^{-1} (Y-\mu_{imx_j})]}$$

Where $\beta_{ij} = \dfrac{\grave{n}_{ij}\,!}{x_{imj}\,!}$; $i = 1,2, \ldots w;\ m = 1,2 \ldots \ell$, and

$X_j = 0, 1, 2 \ldots \grave{n}_j$ (2.7)

let,

$$\beta_{i\nu mx_j} = \Sigma_{mx_j}^{-1} (\mu_{imx_j} - \mu_{\nu mx_j}) \qquad (2.8)$$

$$C_{i\nu mx_j} = -\tfrac{1}{2}(\mu_{imx_j} - \mu_{\nu mx_j})' \Sigma_m^{-1} (\mu_{imx_j} + \mu_{\nu mx_j}) \qquad (2.9)$$

and

$$r_{i\nu mx_j} = \log \frac{\alpha_i\, \beta_{ij}\, \pi_i\, P_{im}\, \phi_{imx_j}^{x_{imj}}}{\alpha_\nu\, \beta_{\nu j}\, \pi_\nu\, P_{\nu m}\, \phi_{\nu mx_j}^{x_{\nu mj}}} \qquad (2.10)$$

Then (2.7) can be simplified as

$$P(H_1/\zeta) = \frac{\prod_{j=1}^{t} \exp\left(\Upsilon' B_{1\nu mx_j} + C_{1\nu mx_j} + \gamma_{1\nu mx_j}\right)}{1+\sum_{1\neq\nu}^{w}\prod_{j=1}^{t} \exp\left(\Upsilon' B_{1\nu mx_j} + C_{1\nu mx_j} + \gamma_{1\nu mx_j}\right)} \qquad (2.11)$$

and hence,

$$P(H_1/\zeta) = \frac{\exp\left(\sum_{j=1}^{t}\left(B_{1\nu mx_j} + C_{1\nu mx_j} + \gamma_{1\nu mx_j}\right)\right)}{1+\sum_{1\neq\nu}^{w} \exp\left(\sum_{j=1}^{t} B_{1\nu mx} + C_{1\nu mx_j} + \gamma_{1\nu mx_j}\right)} \qquad (2.11)'$$

If we put,

$$\sum_{j=1}^{t}\left(\Upsilon' B_{1\nu mx_j} + C_{1\nu mx_j} + \gamma'_{1\nu mx_j}\right) = \eta_{1mx_j} \qquad (2.12)$$

then (2.11) will be

$$P(H_1/\zeta) = \frac{e^{\eta_{1mx_j}}}{1+\sum_{1\neq 1}^{w} e^{\eta_{1mx_j}}} \qquad (3.13)$$

Thus, η_{1mx_j} is positive when H_1 is more probable, and H_ν is more probable if η_{1mx_j} is negative. Hence, the allocation rule is to (classify an observation ζ (Z, X, Y) into H_1 if $\eta_{1mx_j} > 0$, otherwise it is classified into H_ν ($1 \neq \nu = 1, 2, \ldots, W$). (Day and Kerridge 1967).

(3) PROBABILITY OF MISCLASSIFCATION

It should be noted that practical use of the allocation rule is likely to result in some mistakes in classification. If the relative costs of these mistakes (misclassification) can be estimated, the rule should take them into account.

Let $C(\nu/i)$ be the cost incurred when an observation which actually belongs to group i is classified as belonging to group ν.

Now, given ζ is from H_1, the conditional distribution of ζ given X and Z is a multivariate distribution with mean μ_{1mx_j} and dispersion matrix,

$$D^2_{mx_j} = (\mu_{1mx_j} - \mu_{\nu mx_j})' \Sigma^{-1}_{mx_j} (\mu_{1mx_j} - \mu_{1mx_j}) \qquad (3.1)$$

Which is the Mahalanobis squared distance between H_1 and H_ν, conditional on the observation falling in multinomial cell m for the discrete variable X_j.

Note that the overall probability of misclassification from H_1 is the sum of the probabilities of misclassification for each multinomial cell of H_1 weighted by the probability of X_j in this cell.

Now, let

$$A = \left(\frac{\pi_1 C(\nu/1)}{\pi_\nu C(1/\nu)}\right) \qquad (3.2)$$

According to Chang and Afifi (1974) and Krzanowski (1975, 1980), the probability of misclassifying an observation from H_1 into H_ν is

$$P(\nu/1) = \sum_m^{\ell} \sum_x^{n} \prod_{j=1}^{p} \frac{n_{1m}!}{x_{1mj}!} P_{1m} \phi_{1mx_j}^{x_{1mj}} F[(\log A - \tfrac{1}{2} D^2_{mx_j})/D_{mx_j}]$$

and $\qquad (3.3)$

$$P(1/\nu) = \sum_m^{\ell} \sum_x^{n} \prod_{j=1}^{p} \frac{n_{\nu m}!}{x_{\nu mj}!} P_{\nu m} \phi_{\nu mx_j}^{x_{\nu mj}} F[(\log A^{-1} - \tfrac{1}{2} D^2_{mx_j})/D_{mx_j}]$$

When F(u) is the cumulative standard normal distribution function. $\qquad (3.4)$