International Conference on Science, Technology, Education, Arts, Management and Social Sciences
iSTEAMS Research Nexus Conference, Afe Babalola University, Ado-Ekiti, Nigeria, May, 2014.

A Statistical Discriminant Analysis and Classification of Eggs Produced by Poultry Birds Fed with Organic and Inorganic Copper Salt

Ayanlowo E.A & Efuwape B.T
Department of mathematical Sciences
Olabisi Onabanjo University
Ago-Iwoye, Nigeria
Corresponding Author E-Mail: loli4lov@yahoo.com

ABSTRACT

This thesis, discuss the theory of discriminant analysis and how it can be used to establish differences between eggs produced by poultry birds fed with organic copper salt from those fed with inorganic copper salt. The test on the significance of difference in means of the variables shows a significant difference in means of some variables of interest. These suggest that the two groups can be separated via these significant variables, using the "discriminant model". The discriminant models obtained thus classify accurately the data group sets under study as well as future data. In summary, discriminant analysis and classification procedure is an important and reliable tool in determining the accurate functional model required to separate two or more groups of related number of variables. The model of a discriminant function should follow multivariate normal distribution approach when the data under study is quantitative (measured). The principle of cross validation reveals that the cholesterol content of eggs classified under the organic copper salt data are lower compared to that of inorganic copper salt data. The cholesterol content which is majorly determined by the variable x_4 and x_5 shows a classification rule given by allocate x to organic copper salt data section if $x_5 < 160$mg/egg and to inorganic copper salt data section if otherwise. Hence, we are able to separate eggs with high cholesterol content from those with low cholesterol content. This will successively reduce to the barest minimum heart problems caused by consumption of high cholesterol.

Keywords: Statistical Discriminant Analysis, Classification, Eggs, Poultry Birds, Fed Organic, Inorganic Copper Salt

1. INTRODUCTION

1.1 Discrimination And Classification

Discrimination refers to the use of discriminant functions to maximize the separation between groups of available individuals. Classification on the other hand, refers to the rules which are required to minimise the misclassification rate over all possible future allocations.The best classification rules is that which leads to the smallest probability of misclassifying variables and in statistical term this is equivalent to the rules that leads to the smallest error rate of all future allocations of individual samples. The theoretical basis underlying discrimination and classification is base on Baye's minimum error rule. However, a more convenient formulation of this rule is obtained by applying Baye's theorem.

$$P(w_1/x) = P(x\,w_1)P(w_1) \Big/ P(x)$$

The theorem implies that variable X is said to be an object that belongs to the class *w* of the population Ψ provided that the probability that the object x comes from class w is greater that its probability for any other classes. For two classes the Bayes minimum error classification rule can be expressed in the convenient likelihood ration form

$$\frac{P(x\,w_1)}{P(x)} \;><\; \frac{P(w_2)}{P(w_1)} \;\rightarrow\; X_\varepsilon \begin{cases} \Omega_1 \text{ or } \Omega_1 \\ \Omega_2 \text{ or } \Omega_2 \end{cases}$$

$$\text{Or } h(x) \;><\; \text{Constant} \;\rightarrow\; X_\varepsilon \begin{cases} \Omega_1 \text{ or } \Omega_1 \\ \Omega_2 \text{ or } \Omega_2 \end{cases}$$

Where $h(x) = \frac{P(x\,w_1)}{P(x\,w_2)}$ is the discriminant function required for accurate separation.

Ω_1, Ω_2 is the population for classes w_1 and w_2 respectively. For simplicity the constant usually denoted as *k* is assumed to take a value of unity.

1.1.1 Discriminant Function Analysis

Discriminant function analysis is concerned with the problem of seeing whether it is possible to separate different groups on the basis of he available measurements. This could be sued, for example, to see now well surviving and non-surviving sparrows can be separated using their body measurements as the variable of separation. However, we can as well be concerned with how secure from different epochs can be separated, again using size measurements as variable of separation. Like principal component analysis, discriminant function analysis is based on the idea of finding suitable model which is a combination of the original characteristic variables. In situations where more than one variables are required to achieve the necessary classification and form a reliable comprehensive model, the use of matrix algebra in describing the format of presenting data for a discriminant function analysis is inevitable. This in general for a well designed discriminant function analysis, there will be *m* random samples, from different groups of sizes n_1, n_2, …, n_M and values will be available for P variables x_1, x_2, …, x_p for each sample member. Hence the data for a discriminant function analysis takes the form below up to M^m number of groups.

P variables

Sample of individual	X_1	X_2	…	X_p	
1	X_{111}	X_{121}	…	X_{1P1}	Group 1
2	X_{211}	X_{221}	…	X_{2P1}	
.	.	.	.	.	
.	.	.	.	.	
.	.	.	.	.	
n_1	Xn_111	Xn_121	Xm…	Xn_1P1	
1	X_{112}	X_{121}	…	X_{1P2}	Group 1
2	X_{212}	X_{221}	…	X_{2P2}	
.	.	.	.	.	
.	.	.	.	.	
.	.	.	.	.	
n_2	Xn_212	Xn_222	…	Xn_2P_2	
.	.	.	.	.	
.	.	.	.	.	
.	.	.	.	.	
1	X_{11m}	X_{12m}	…	X_{1Pm}	Group m
2	X_{21m}	X_{22m}	…	X_{2Pm}	
.	.	.	.	.	
.	.	.	.	.	
.	.	.	.	.	
n_m	Xn_m/m	Xn_m2m	…	Xn_mPm	

The main advantage of discriminant function analysis is that the data to be analysed do not need to be standardised to have zero means and unit variances as it is the case for principal component and factor analysis. This is because the outcome of a discriminant function analysis is not affected in any important way by the sealing individual variables.

1.1.2 DISCRIMINANT MODEL

Discriminant function as earlier been denoted by *h(x)* leading to the classification rule.

$$h(x) \quad >< \quad k \quad \rightarrow \quad X_{\varepsilon}\begin{cases}\psi_1 \\ \Psi_2\end{cases} \text{ where } k \text{ is a constant}$$

The model of discriminant function *h(x)* usually assumed a linear or a quadratic form depending on the test hypothesis carried out on the matrices of the means and covariances of the different groups under investigation. To determine if the groups can be taken to be from the same population or not we take the following steps.

If the covariance matrices is denoted by $\sum$ setting up two different hypothesis test as:

Null hypothesis test H_0 : $\sum_1 = \sum_2 = \sum$

Alternative hypothesis H_1 : $\sum_1 \neq \sum_2$

Then we conclude that if the null hypothesis test is true that the covariances of the different groups are much more the same and can be combined as a pooled variance then the discriminant function *h(x)* will assume a linear model

$h(x) = V'x + a_0$

Or $h(x) = \sum_{i=0} a_0 x_i$ Where $x_0 = 1$

The practical benefits of obtaining this assumption are that the discriminant function and the classification (allocation) rule became very simple and linear. Linear discriminant function is usually denoted by *L(X)*.

On the other hand, if the alternative hypothesis test is true, in this case the covariance matrix of different groups are distinct and different from each other, then the discriminant function *h(x)* will be a quadratic model type discriminant function denoted by Q*(x)*.
The quadratic discriminant function model *Q(x)* usually contains a quadratic matrix as follow *Q(x) = X'Ax*

The allocation rule of both linear and quadratic model discriminant function is given as

$$L(x) \quad >< \quad k \quad \rightarrow \quad X_{\varepsilon}\begin{cases}\Omega_1 \\ \Omega_2\end{cases}$$

$$Q(X) \quad >< \quad k \quad \rightarrow \quad X_{\varepsilon}\begin{cases}\Omega_1 \\ \Omega_2\end{cases}$$

For simplicity take $k = 1$

1.1.3 GENERAL DISCRIMINANT MODEL

The most general form of the model is to assume that Ψ, is a multivariate normal population with mean μ, and variance-covariance (dispersion) matrix i, for i = 1, 2 two group case.
The probability density function for P-variables ($x_1, x_2, x_3, \ldots, x_p$) of ith population is given by

$f_i(x) = (2\lambda)^{-p2} \left[\sum i\right]^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(x-\mu i)^{-1}\sum i^{-1}(X-\mu i)\right]$

for $i = 1.2$

The allocation rule given by

$$\frac{f_1(x)}{f_2(x)} \quad >< \quad k \quad \rightarrow \quad X_\varepsilon \begin{cases} \Omega_1 \\ \Omega_2 \end{cases}$$

Now takes an algebra form of

$$\frac{f_1(x)}{f_2(x)} = \left[\sum 2\right]^{-\frac{1}{2}} \left[\sum 1\right]^{-\frac{1}{2}} \exp\left[-\frac{1}{2}x^1\sum 2^{-1} - \sum 2^{-1}(X-\mu_i 2)\right]$$

Which is the generalised discriminant model for two group case assuming a muitivariate normal distribution.

The allocation rule still implies that allocate to x. if $\frac{f_1(x)}{f_2(x)}$ > k or else allocate x to Ψ_2.

1.1.4 GENERALISED DISCRIMINANT MODEL SIMPLIFICATION

The two group case generalised model obtained earlier contains a quadratic form $X(\sum 2^{-1} - \sum 2^{-1})x$. Thus it can be reffered to as quadratic discriminant function *Q(x)*.

For simplification to obtain the linear discriminant function estimate we substitute $\Sigma_1 = \Sigma_2 = \Sigma$ which turn the Quadratic form term to *zero* with he linear discriminant function *L(X)* obtained as *L(X)*.

1.2 COPPER AVAILABILITY AND ROLES

Copper is a trace element that is required in small amount in the diet. The amount required by hers, is derived ogten times from the conventional feedstuffs such as corn and bye products, wheat, yeast and most importantly mineral premix.

However, copper availability in the feed may be significantly reduced by some antagonistic elements in the diet such as zinc, rnageed and ochme. Thus supplementation often times because important. Copper can be supplemented using either organic source or in-organic source, e.g. copper proteinate, coppersulphate tribasic soppet chloride e.t.c.

However, studies have shown that absorption rate of copper in organic form was higher than in-organic form was higher than in-organic forms, other studies indicated that copper complexes were no more effective than copper surlphate (an in-organic form) in improving copper status of chiker.

Copper is required for the activity of enzymes associated with Iron metabolism.

It acts as a processor by which iron is metabolised for normal red blood cells formations. Copper is also responsible for elastic and collagen formation, Malaren production and the integrity of the central nervous system.

Copper is required for normal bone formation and for normal elastic formation in the aorta and remainder of cardio vascular system.

Moreover, copper is needed for the production of oxidating enzymes such as tyrosine's, monoamise oxidase ceruloplasmin gaiactuse oxidase and cytochrome oxidase.

1.2.1 DESCRIPTION OF THE STATE OF ORGANIC AND IN-ORGANIC SALT

ORGANIC COPPER SALT

Copper proteinate is a good source of organic copper. It is greenish in colour and in powdery form. A mineral proteinate is a mineral complex formed by reacting a mineral salt with a specifically prepared mixture of amino acid and small peptide.

IN-ORGANIC COPPER SALT

Copper proteinate is a source of inorganic copper. It is however bluish in colour and also in powdery form. It is slightly hygroscopic. It is formed mostly by chemical reaction in the laboratory.

1.2.2 EFFECT OF ORGANIC AND INORGANIC COPPER SALT

However, scientific research carried out by Bakalli et al 1995, National Research council 1994, Ghiselli 1992, Pesti and Bakalii 1996, Konjufca et al 1997 an so non has extensively show that copper slat (organic and inorganic) has been found to reduce lipid metabolism, with more reduction with naturally ovvuring organic and inorganic copper slat is it's ability in reducing he cholesterol content of egg.

However, scientific research has shown that organic copper salt has the ability of reducing the cholesterol content beyond the harmful level.

Research has also shown that the higher the concentration of copper feed utilixed the lower the cholesterol content and that the relationship between copper feed (organic or inorganic concentration and cholesterol content is al inear reduction association.

1.3 SCOPE OF STUDY

The scope of the project is based on the classification of egg samples obtained from poultry birds into classes of egg with low or high cholesterol content using discriminant function analysis.

The data required for the project work will be taken from the poultry unit of livestock production of Ayokunle farm in Ijebu-Ode, Ogun state.

1.4 AIMS AND OBJECTIVES OF THE STUDY

The aim of this project work is to formulate a reliable discriminant model (linear or quadratic) and a classification rule which can predict with high degree of accuracy the poultry products (eggs) whose cholesterol has considerably reduced beyond the harmful level.

In mathematical term, the aim of the project work is to find a function mapping objects to an index set consisting of c ass identifiers.

The objectives of this research include the following:

(1) To determine a discriminant function analysis that can enhance maximum and accurate separation of variables in different groups.
(2) To construct a classification rule required to minimise the error of misclassifying objects.
(3) To check if there are significant differences in the mean values of the variables in the different groups.
(4) To fin out if there are significant differences in the sample standard deviations for the variables and if so, do the differences reflect gradual changes with time.
(5) To determine if there variables characteristics are related to each other or not.

2. RELATED LITERATURE

2.1 INTRODUCTION

Statisticians all over the world have contributed in ensuring an accurate classification of objects that look similar but differ in one way or the other using discriminant analysis approach.The oldest research in discrimination and classification was proposed in 1936 by Fisher. Since that time lots of researchers have continuously being carried out to improve further on the research work.

2.1.1 Fishers Linear Discriminant Function

In contrast to the probabilistic approach of allocating variables given by allocate x to π_1 if $f_1(x)/f_2(x) > 1$ and to x_2 if $f_1(x)/f_1(x) \leq 1$. Fishers in 1936 tackled discrimination from a purely data-based stand point. He supposed that given an independent random sample, of sizes n_1 and n_2 respectively, from each of two P-variate populations and that a method of best is distinguishing between these samples was required. Fishers further made an assumption that the dispersion matrices in these two populations were equal otherwise, the population were completely unspecified.

With this assumption, the data can be summarised by computing the sample mean vector $\overline{X}^{I}$ and $\overline{X}^{II}$ and the problem within sample covariance matrix S. Fisher then looked for the linear combination w = ax of responses that gave maximum separation of the group means, when measured relatively within S - group variance of the data. This linear combination he found by maximising $[a(\overline{X}^{I} - \overline{X}^{II})\ S^{-1}X$. Fishers then further proposed an allocation rule for the above equation by allocating x to x_1 if $(\overline{X}^{I} - \overline{X}^{II})^{I}\ S^{-1}X$ is generally known as Fisher's Linear Discriminant Function (FLDF).

The advantage officers' discriminant function is that the function requires no probability density function. Thus, fishers' linear discriminant function should provide a useful tool for discrimination under wide distributional 'conditions. Furthermore, fishers in 1938 went to show that if dummy dependent variable to is defined to have one value for each individual in the sample $7t_1$ and another (different) value for each individual in the sample from x_2, then regression equation for some constant. Thus, the fisher's linear discriminant function that best predicts group membership of a sample individual, Since Fisher's LDF is a linear function then its best applied on any data that is already Normal because linear and normal are interchangeable.

Validation Of Fisher's Linear Discrimination Function

Many statisticians criticised the discriminant mode of fisher's through its application on some circumstances, which shows an unsatisfactory performance. The statistician that contributes to fisher's validation and subsequently its improvement include the following. Lachenbruch et al. in 1973 considered continuous but non-normal data on the fisher's discriminant function model. The results shows that fisher's linear discriminant function can be greatly affected by non-normality, in the sense that the error rates can be in balance, with those for one population being greater than the optimum values, while those for the other population are less than the optimum values, and the sum of the optimum values, and the sum of the two error rates increases for some distributions. Lachenbruch also observed that the fishers' discriminant function is accurate and easy to compute for situations in which normality can be achieved.

Hence, Lachenbruch recommended transforming ail data under study to appropriate normality before using the fisher's inear discriminant function. Moore also studies the effectiveness and accuracy of fisher's linear discriminant function in 1973 by applying the fisher's model to binary

data that is to data's coded with values 0's and 1's. Moore then observed that the fisher's model performed badly for certain situations of the coded values.
Moore further established that if the log-livelihood ratio for the two population is plotted against the number of variables having the value 1, then in some populations this log-likelihood ratio will not increase rnotonically and is said to undergo a 'reversal'. Fisher's LDF, on the other hand, may increase monoionical y for this population; if it does increase monotonically then it will not be able to follow such 'reversals'. Moore then concluded that fisher's model in such circumstances will have a worse performance. Krazanwoski in 1977 considered the application of fisher's linear discriminant function to mixtures of binary and continuous data, and also found circumstances in which it performed poorly. Broadly speaking, he stated that positively correlated binary variables or sign reversal sin the binary/continuous correlations from one population to the other are warning signs that the LDF may do well. Krazanwoski then suggested the use of location model approach has a preferable method of solving the problem of fisher's discriminant function. In conclusion of the fisher's model validation, the message is clear from all the research studies that it is absolutely wrong and ambiguous to trust on the sole of the linear discriminant function (LDF) if the data show gross departure from multivariate normality.

Logistic Discrimination
Cox (1966), Day and Kerridge (1967) started from a probabilistic methods given as allocate x to X_1 if $f_1(x)/f_2(x) > 1$ and hence otherwise to π_2 which requires parametric specification of each density function $f_1(x)$ followed by estimation of $f_1(x)/f_2(x)$ from the training set and evaluation of this ratio for the individual x to be allocated. An alternative basis for discrimination and classification proposed by these authors is a parametric specification of the posterior probabilities $q(\pi_1/x)$ and $q(x_2/x)$. The allocation rule provided above now becomes allocated x to π_1 if $f_1(x)/f_2(x) > q_2q_1$ and to π_2.

If $f_1(x)/f_2(x) = q_2/q_1$ provided that $q(\pi_1 / x_0) = q_1 f_1(x_0) \Big/ q_1 f_1(x_0) + q_2 f_2(x_0)$

For an observed vector x_0
Cox, et al. them propose a possible model for the possible model for the posterior probabilities called the logistic model given by

$$q(\pi_1/x) = \exp(d_0 + dx) \Big/ 1 + \exp(d_0 + dx)$$

$$q(\pi_2/x) = 1 \Big/ 1 + \exp(d_0 + ax)$$

With this mode, the posterior log adds - ratio is a linear function of the observed variables that is $\text{Log}_e\ [q(\pi_1/x)/q(\pi_2/x)] = d_0 + d$ be and hence application of the allocation rule based linear discriminant function. The model only requires the parameters do and a for its application.
In his own contribution J.A. Anderson (1972) mentioned the need to include the probability density function $f_1(x)$ and estimation of many more parameters over all the variables. The various parameters of interest include means, variances, covariance etc.

Moreover, J.A. Anderson pointed out that the Log odds-ratio is linear in x for a range of different assumptions about the $f_1(x)$. For instance he said for every multivariate normal density with common dispersion parameters, for independent binary variables, for multivariate discrete distributions all follow the log-linear model with the same interaction in each population or group.
Hence, he concluded that logistic model will be optimal under a wide range of data Pes. J.A. Anderson suggested the estimation of parameters through the method of iterative ewtan-Raphsan procedures.

2.3.1 Recommendation of Logistic Discrimination

In view of the close affinity between the logistic and the fishers' discriminant function, the performance of logistic discrimination has been studied in relation to linear discriminant analysis by a number of authors. The general consensus is that logistic discrimination is to be referred when the distributions are clearly non-normal or the dispersion matrices are clearly unequal. Any deviation of the dispersion matrices from equality, the results of the methods (logistic and fisher's model) are likely to be very similar.

2.4 Distance-Based Discrimination

Penrose Distance

Penrose in 1953 proposed a distance measure that can be required in partitioning a group of class can be required in partitioning a group or class. He considered two multivariate populations when information is available on the means, variances and covariance of the populations. Penrose noted it that supposed a multi-population are available and the multivariate distributions in these populations are known for P variant X_1, X_2, …, X_p. Let the means of variable X_k in the i^{th} population be μ_{ki} and assume that the variance of x_k is the same value, v_k in a population. The measure proposed by Penrose is given by;

$$P_{ij} = \sum (X_{ki} - \mu_{kj})^2 \Big/ PV_k$$

The main disadvantage of Penrose measure is that it does not take into account the correlations between the P variables. This means that when two variables are measuring essentially the same thing and are highly correlated, the variable still behave differently and contribute about the same amount to the population distance such that the result becomes independent of all the other variables.

Mahalanabis Distance

Mahalanabis worked on distance- based determination and proposed a good distance measure for proper discrimination and class function. The measure does take into account correlations between variables under study. Mahalanabis pre-supposes that the two training sets in any discrimination problem may be thought of as two swarms of points in p-dimensional space. The greater the difference between the population x_1 and x_2 the greater will be separation between the two swarms. An individual x to be allocated tone of x_1 and x_2 may then be thought of as a singe point in this space and an intuitively attractive procedure would be required to allocate x to the population it accurately belong.

Mahalanabis approach requires a definition of distance between the single observation x and each training set sample Mahalanabis quantity D_i^2 is given by $D_i^2 = (x - x^i)\, S^{-1}\, (x - x^1)$ where x^1 is the mean of the i^{th} trainings et (i = 1, 2) and S is the covariance matrix poorled within the training set. Allocation of x would then be to population π_1 if $D_1^2 < D_2^2$ and to x_2 if D_1^2 D_2^2. Mahalanabis is thus a measure of how far the observation x is from the centre of the distributions of all values, taking into account all the variables being considered. In conclusion, extensive criticism has shown that the Mahalanabis distance generate functions that are monotonic in behaviour when $f_1(x)$ and $f_2(x)$ are multivariate normal with cannon dispersion motive. An important and useful result is that if the population being considered is multivariate normally distributed, then the values of D^2 will follow a chi-squared distribution with P variables degrees of freedom. A significantly large value of Di^2 mean that the estimated values of population means, variances and covariance must be unbiased.

2.5 General Consensus On Distance-Based Discrimination

In principle the Mahalanabis distance is superior to the Penrose distance because it uses information on covariances.
However, the advantage of Mahalanabis distance is only present when covariances are accurately known. That is when covariances can only be estimated with dom it is probably best to use the simpler Penrose measure.

3. METHODOLOGY AND DATA ANALYSIS

3.1 Source Of Data

3.1.1 Location Of Research Station

This research was carried out in the research unit of the livestock farms, a production and research farm at Federal University of Agriculture, Abeokuta, Ogun State. All facilities used were supplied by the Poultry Management and Technical committee of the livestock Production Department, Federal University of Agriculture, Abeokuta, Ogun State Nigeria which lies in the tropical climate with ambient temperature of 34^0C and mean annual rainfall of 1100mm.

3.1.2 Experimental Birds And Management

Ninety-six (old) layers of black harce strain were obtained from POMTEC. The binds were randomly allotted to six treatments of four replicates where each replicate contains four birds. The ninety-six birds were randomly allotted at two (2) birds per battery in a battery cage system. The battery cages were placed in an open roof housing system, with dwarf wall of about 1.5m high. The open and were screened up with wine mesh to keep out wild birds.
Asbestos were used as roofing materials

3.1.3 FEEDING TRIALS AND MANAGEMENT PROCEDURES

Routine management in intensive system was carried out on the birds. Feed and water were given early in the morning of everyday of the experiment. The dietary treatment consisted of six treatment of organic copper salt and another treatment of inorganic copper salt. The concentration of two copper feeds is maintained at 150mg/kg. The naturally occurring organic copper feed is called copper proteinate while that of inorganic copper feed is copper sulphate. The variables of characteristics that are of interest one egg weight (x_1), yolk wet weight (x_2), yolk dry weight (x_3), yolk cholesterol (mg/g) (x_4). cholesterol/egg (mg/egg) (x_5) and yolk ratio (x_6).

3.1.4 Parameters Measured

(a) Live Weight and Feed Intake

Feed consumption of each replicate was recorded on weekly basis. It was measured by subtracting the weight of the residual feed from that of the feed offered to the birds at the beginning of the week. The birds live weights were also taken on weekly basis.

(b) Cholesterol Content

Daily egg production was kept on replicate basis, which means, daily egg production were recorded on replicate basis. Egg samplings were done very two weeks for cholesterol analysis. The total number of 48eggs (2 per replicate) were collected per sampling for a period of 6 weeks. The egg weight, wet yolk weight, dry yolk weight and dry matter percentage of yolk were recorded. Sampled eggs are usually weighed and oiled within 24 hours of lay. The eggs were then allowed to cool at room temperature after 8 minutes of boiling. Yolk and albumen were separated and weighed. Yolks were oven dried at 77^0C to constant weight and stored in the desiccators until the time of analysis. The yolk lipid was extracted with chloroform-methanol mixture. The resulting mixture was then immersed in a boiling water bath, which was removed and shaken gently immediately the solvent started to boil. It was then cooled and centrifuged for 3 minutes. The supernatant fluid was decanted into another test-tube for yolk cholesterol analysis.

3.2 Analytical Methodology

3.2.1 Discrimination And Classification Format

In the general case, there will be m random samples from different groups of sizes n_1, n_2, …, n_m independent observations and values of measurement will be available for all the P variable characteristics for each sample member. Thus, the format of the data for the purpose under study is presented as follows.

P VALUES

Sample of individual	X_1	X_2	…	X_p	
1	X_{111}	X_{121}	…	X_{1P1}	Group 1
2	X_{211}	X_{221}	…	X_{2P1}	
.	.	.	.	.	
.	.	.	.	.	
.	.	.	.	.	
n_1	Xn_111	Xn_121	Xm…	Xn_1P1	
1	X_{112}	X_{121}	…	X_{1P2}	Group 1
2	X_{212}	X_{221}	…	X_{2P2}	
.	.	.	.	.	
.	.	.	.	.	
.	.	.	.	.	
n_2	Xn_212	Xn_222	…	Xn_2P_2	
.	.	.	.	.	
.	.	.	.	.	
.	.	.	.	.	
1	X_{11m}	X_{12m}	…	X_{1Pm}	Group m
2	X_{21m}	X_{22m}	…	X_{2Pm}	
.	.	.	.	.	
.	.	.	.	.	
.	.	.	.	.	
n_m	Xn_m/m	Xn_m2m	…	$X_{nm}Pm$	

But for the purpose of the project work, work, we shall be restricted to two group cases alone (organic and in-organic group cases). The original data for the project study is present at the appendix.

It is obvious that the format presentation is a true representation of matrices algebra. Therefore it is paramount to have certain minimal knowledge of this area of mathematics.

3.2.2 Methodology Theory

In the un-variate case of normal distribution, the variables are represented by a mean population estimated alongside with it's variance denoted by X - N (N, θ^2). In the discrimination and classification procedure, since the variables of interest are many (P varaite where P > 1) then they can be summarised by using the mean vectors and the variance -covariance matrices. Considering the P variates, x_1, x_2, …, x_p, taking the expectations of the variables we have $E(x_1)$, $E(x_2)$, … , $E(x_p)$ which generate a sample estimate of the population vector means.

Population mean vectors is given as [N_1, N_2, …, N_p] while sample vector means is [$\overline{X}_1, \overline{X}_2$, … $\overline{X}_p$].

Also, for the variance - covariance matrices required we have the variances given by $V(x_1)$, $V(x_2)$, … $V(x_p)$ occupying the diagonal spots of a matrices algebra and the covanances between different variants occupying the off-diagonal spots.

To represent the variance - covanance matric in a diagrammatic format we have:

	X_1	X	…	X_p
X_1	$V(X_1)$	$Cov(X_2, X_1)$	…	$Cov(X_p, X_1)$
X_2	$V(X_1, X_2)$	$V(X_2)$	…	$Cov(X_p, X_2)$
.	.	.	…	.
.	.	.	…	.
.	.	.	…	.
X_p	$V(X_1, X_p)$	$V(X_2, X_p)$	…	$V(X_p)$

The above matrices demonstrate a P x P dimensional matrix of rank r. Let refers to the variance - covariance matrices above as dispersion matrix denoted by Σ. Then for the two group cases under study we have the following notations.

	Group 1	Group 2
Population Vector Means	$X_1, X_2, \ldots Xp$	$X_1, X_2, \ldots Xp$
Sample Vector Means	$(\mu, \mu 2, \ldots, \mu p) = \overline{X}_1$	$(\mu, \mu 2, \ldots, \mu p) = \overline{X}_1$
Variance Covariance Matrix	Σ_1	Σ_1
Population Distribution	$\mu(\mu_1, \Sigma_1)$	$\mu(\mu_1, \Sigma_1)$
	$\mu(X_1, \Sigma_1)$	$\mu(X_1, \Sigma_1)$
Sample Distribution Representation		

The unbiased estimates of the population parameters μ, μ, Σ^1, Σ^2, are denoted and calculated as follows:

$$X_1 = \sum X_1 \Big/ n_1$$

Where i = 1, 2, …, p and calculated separate for the different groups. In the multivariate (P variable) cases X_1, X_2, …, X_P with mean K and variance - covariance matrix ∑ then the probability density function is given by:

$$f(x) = 1/C|\sum| \, ½ \, (2\pi)^{p/2})e^{-½} \, (X - \mu)\Sigma^{-1}(x - \mu)$$

Where the splitting of $(x - \mu)^2$ to $(x - \mu)$ $(x - \mu)$ is in line with the theory of matrices quadratic variables.

The necessary comments and assumptions based on multivariate normal distribution to be used as model are given bellow;

1) The variance - covariance matrix ∑ must be non-singular that is $|\Sigma| \neq 0$

2) The transformation $Y = (x - \mu)/ |\Sigma|^{1/2}$ is used to standardise the P-variates as $Z = (x - \mu)/\sigma$ is used in the uni-variate case.
3) The variance-covariance-covariance matrix is symmetric. $Cov(x_i, x_j)$ - $COV (x_j, X_i)$.
4) The distribution is usually given by $X \sim NP [\mu, \Sigma]$.

Discriminant Model

The discriminant model of the multivariate normal distribution function required can be linear or quadratic.

To determine the best model, we desired to test the null hypothesis H_0: $\Sigma_1 = \Sigma_2 = \Sigma$ against the alternative H_1: $\Sigma_1 \neq \Sigma_2$.

For the two group cases if $\Sigma_1 = 1/(n_1 - 1)\ \Sigma(x_i - \overline{X})\ (x_i - x)$ and $\Sigma_2 = 1/(n_2-1)\ \Sigma(x_i - \overline{X})(x_i - x)$ then

$\Sigma = 1/(n_1 + n_2 - 2)\ (c_1 + c_2)$ where $c_1 = \Sigma(x_i - x)\ (x_i - x)$ and $c_2 = \Sigma(x_i - x)\ (x_i - x)$.

Test statistic A Given by
$-2 \text{ In } \lambda = (n_1 + n_2 - 2) \text{ In } |\Sigma| - (n_1 - 1) \text{ In } |\Sigma|_2$ is utilised. The quantity A has a chi-squared distribution with $^1/_2\ (P + 1)\ (P)$ degrees of freedom. If the null hypothesis H_0 is accepted then we assume a linear discriminant model but if otherwise a quadratic discriminant model is assumed- The generalised discriminant function has a quadratic discriminant model denoted by $\hat{Q}(x)$ and given as

$\hat{Q}(x) = (f_1(x)/f_2(x)|\ \Sigma|^{-1/2} \exp. [-\frac{1}{2} X^1 (\Sigma_1^{-1} - \Sigma_2^{-1})X - 2x^1 (\Sigma_1^{-1} \mu^I - \Sigma_2^{-1} \mu^{II}) +$
$(\mu^I)' \Sigma_1^{-1} \mu^I - (\mu^{II})' \Sigma_2^{-1} \mu^{II}]$ while the discriminant model assumed a linear form given by $L(x) = (\mu^I - \mu^{II})\Sigma^{-1}[x - \frac{1}{2} (\mu^I + \mu^{II})]$ where L(x) is a linear discriminant model.

Classification Rule

In both cases, the allocation rule being suggested is of the form: Allocate x to π_1, if $f_1(x)/f_2(x) > k$ and to π_2 if $f_1(x)/f_2(x) \leq k$ where $k > 1$. The value of k should depend on the prior probabilities of the population π_1 and π_2.. First we will assume that q_1 is the prior probability that an observed value of x is from π_1 and that q_2 is the prior probability that the new observed variable x cause from populations π_2 (with $q_1 + q_2 = 1$). Secondly we will assume that c(½) is the cost incurred whenever an individual from π_2 is incorrectly allocated to π_1 and that c(2/1) is the cost of misallocating an individual from π_1 and π_1. Using the cost function of misallocation then we assume that the best allocation rule is the one that yields minimum expected cost due to misallocation.

By minimization principle we have
$C(^1/_2)q_2f_2(x) - C(2/1)q_1fi(x) < 0$ for which

$$f_1(x)/f_2(x) > C(1/2)q_2/C(2/1)q_1$$

For simplicity we set $c(^1/_2) = c(2/1)$ this we have

$$f_1(x)/f_2(x) > q_1/q_2$$

Where $k = q_2/q_1$ and is called the allocation rule such that any observed individual x is allowed to π_1.

If $fi(x)/f_2(x) > (q_2/q_1)$ and to population π_2 if $f_1(x)/f_2(x) \leq (q_2/q_1)$
Suppose, as before that $f_1(x)$, $f_2(x)$ are the probability densities of x in π_1 and π_2 respectively then the posterior probability that an individual observed vector x_0 comes:

$$q(\pi_1/x_0) = q_i f_i(x_0) \Big/ q_1 f_1(x_0) + q_2 f_2(x_0)$$

4. FOUR: DATA ANALYSIS

4.1 Eggs Produced Using Organic Copper Salt

The sample vectors means for organic copper salt group $\overline{X}_1$ is given as a row matrix below.

$\overline{X}_1$ = [59.09979167, 14.295625, 8.286875, 10.99479, 131.5971, 54.85604]

Each are obtained from the expression;

$\overline{X}_1 = (1/n) \sum X_i$

For instance, 59.09979167 is obtained from egg weight column x_1 of the organic copper salt group as 1/48 (56.08 + 56.34 + - - - + 56,34 + 56,08) = 59.09979167.

Also, for x_5 (cholesterol/egg) we have

1/48 (60.73 + 66.03 + - - - + 87.16 + 81.86) = 131.5671

The variance - covariance matrix of the organic copper salt $\sum_1$ is given by the 6 X 6 matrix below:

$$\Sigma_1 = \begin{pmatrix} 0.979167 & 0.9 & 0.912423 & 0.83478 & 0.920279 & 0.954481 \\ 0.9 & 0.979167 & 0.892368 & 0.751735 & 0.879763 & 0.90574 \\ 0.912423 & 0.892368 & 0.979162 & 0.657668 & 0.953505 & 0.93581 \\ 0.83478 & 0.751735 & 0.657668 & 0.979165 & 0.670376 & 0.786647 \\ 0.920279 & 0.879763 & 0.953505 & 0.670376 & 0.979167 & 0.940202 \\ 0.954481 & 0.90573985 & 0.93581 & 0.786647 & 0.940202 & 0.979168 \end{pmatrix}$$

The diagonal are the variance of each column obtained as follows:

$\text{Var}(x_i) = (1/n - 1) \sum (x_i - \overline{X}_j)^2$

For x_3 (yolk dry weight), we have:
$\text{Var}(x_3) - (18.08 - 8.256875)^2 + (8.08 - 8.286875)^2 + \ldots + (8.14 - 8.286875)^2) = 0.979162$

For x_4 (yolk cholesterol), we have:
$\text{Var}(x,) - (10.58 - 10.99479)^2 + (10.63 - 10.99479)^2 + \ldots + (10.15 - 10.99479)^2/47 = 0.979165$
$\text{Var}(x_1)$, $\text{Var}(x_2)$, $\text{Var}(x_5)$ and $\text{Var}(x_6)$ are calculated in a similar way.

The diagonal elements are the covariance obtained from the expression
$C_{ik} = (1/n - 1) [\sum (x_{ij} - \overline{X}_j) (x_{jk} - \overline{X}_k)]$ where i = k = 1, … 6 and n is the number of observation in each column which ware usual y the same.

In the vector form, we can have it as $C_{ik} = (1/n - 1) [(x_i - \overline{X}_j) (x_k - \overline{X}_k)]$

where i = k = 1, … 9.

For instance, the covariance between variable in columns 2 and 3 can be calculated as follow:

$C_{23} = (1/n - 1)(X_2 - \overline{X}_2)(X_3 - \overline{X}_3)$

$C_{23} = (1/47)[(13.64 - 14.295625)(8.07 - 8.286875) + \ldots + (13.61 - 14.295625)(8.14 - 8.286875) = 0.892368$

Also, the covariance between the variables in columns 4 and 6 is obtained as thus:

$C_{46} = (1/n - 1)(x_4 - \overline{X}_4)(x_6 - \overline{X}_6)$

The covariance of other combinations of i's and k's are calculated in a similar way.
The matrix of diagonals (variances) and the off diagonal elements (covariances) now forms the variance-covariance matrix $\sum_1$.

4.2 Eggs Produced Using Inorganic Copper Salt

The sample vector means for the inorganic copper salt group $\overline{X}_2$ is a row matrix given as $\overline{X}_2$ = [58.35167, 14.7325, 6.89125, 13.72083, 195.7317.46.48875].

Also, the variance - covariance matrix for the inorganic copper salt group has a 6 x 6 matrix given below:

$$\Sigma_1 = \begin{Bmatrix} 0.979167 & 0.954191 & 0.898875 & 0.9495 & 0.941101 & 0.873364 \\ 0.95191 & 0.979168 & 0.916565 & 0.958 & 0.931124 & 0.898105 \\ 0.898875 & 0.916665 & 0.979166 & 0.928443 & 0.82537 & 0.941021 \\ 0\ 949499 & 0.958 & 0.928443 & 0.9791667 & 0.919199 & 0.910511 \\ 0.9411005 & 0.9614 & 0.82537 & 0.919199 & 0.979166 & 0.800959 \\ 0.873644 & 0.898105 & 0.941021 & 0.910511 & 0.800959 & 0.979167 \end{Bmatrix}$$

The calculation required to obtained copper salt group $\overline{X}_2$ and $\sum_2$ are the same as that of the organic copper salt; only that a new data labelled as inorganic copper salt will then be used in the calculation.

4.3 Pooled Variance - Covariance Matrix

The estimate of the pooled variance-covariance matrix by $\sum$ is given as

$$\sum = (n_1 - 1)\sum_1 + (n_2 - 1)\sum_2 \Big/ \mu_1 + n_2 - 2$$

Where $n_1 = n_2 = 48$ and $\sum_1$ and $\sum_2$ are the variance covariance for organic and inorganic copper salt.

The first entry inside $\sum$ is obtained as follows:

(47 (0.657668) + 47 (0.928443)/(48 + 48.2) = 0.793072

The main pooled variance covariance matrix

$\sum$ is thus given as:

$$\Sigma_1 = \begin{matrix} 0.97917 & 0.9271 & 0.906211 & 0.892145 & 0.930591 & 0.91405 \\ 0.9271 & 0.919768 & 0.904525 & 0.85485 & 0.9055 & 0.90192 \\ 0.906211 & 0.904525 & 0.979166 & 0.793072 & 0.889438 & 0.9384 \\ 0.892145 & 0.85485 & 0.793072 & 0.979166 & 0.794758 & \\ 0.848579 & & & & & \\ 0.930691 & 0.9055 & 0.889438 & 0.794788 & 0.979166 & \\ 0.870101 & & & & & \\ 0.91405 & 0.90192 & 0.938438 & 0.794788 & 0.979166 & \\ 0.870101 & & & & & \\ 0.91405 & 0.90192 & 0.9384 & 0.848579 & 0.870101 & \\ 0.979167 & & & & & \end{matrix}$$

4.4 Matrix Determinant

The determinants and the inverse matrices for each of $\sum$ above were determined using Microsoft Excel Software.

The command is given by "=MDE TERM (Array)"

The determinant of above matrices denoted by $|\sum_1|$, $|\sum_2|$ and $|\sum_3|$ are 2.5004 X 10^{-6}m 5.24903 X 10^{-7} and 1.63 X 10^{-6} respectively.

The inverse matrices are obtained with the excel command "MINVERSE (Array)". The result for the inverse of the various matrices are present in the appendix.

4.5 Hypothesis Testing

4.5.1 Test of Significance Of Means Difference

Considering two large samples, that is, sample size n ≥ 30, then given that

	Sample 1	Sample 2
No. of observation	n_1	n_2
Mean	m_1	m_2
Standard Deviation	s_1	s_2
Standard Error	$s_1/\sqrt{n_1}$	$s_2/\sqrt{n_2}$

Then, the estimate of the test statistic says t is given by:

$$t = |m_1 - m_2| / \sqrt{[(s_1^2/n_1) + (s_2^2/n_2)]} \sim |\overline{X}_1 - \overline{X}_2| / [(s_1^2/n_1) + (s_2^2/n_2)]$$

(i) Test of difference in mean of egg weight (x_1)

$t = |59.0997167-58.35166667| / (\sqrt{1.84121215475)^2/48) + (3.5970 + 69335)^2/48)) = 0.748125/(\sqrt{0.34087129}) = 1\ 282371$

(ii) Test of difference in mean of yolk wet weight (x_2)

$t = |14.295625 - 14.7325| / (\sqrt{(0.476850957)^2/48) + (0.834341)^2/48)) --3\ 1496 = 3.1496$

(iii) Test of difference in mean of yolk dry weight (x_3)

$t = |8.286875-6.891251 / (\sqrt{(0.117259)^2/48) + (0.693559)^2/48) =$
$1.395625/(\sqrt{0.010307786}) = 13.74630671$

(iv) Test of difference in mean of yolk cholesterol weight (x_4)

$t = |10.99479 - 13.720831| / (\sqrt{(0.463438)^2/43) + (3.119838)^2/48))}$

$= |-2.7260411| / \sqrt{0.207253415}$

$= |-5.9881| = 5.988$

(v) Test of difference in mean of cholesterol weight (x_5)

$t = |(131.5671 - 195.7317| / (\sqrt{(37.48569)^2/48) + (22.1\ 3494)^2/48))}$

$= |-64.16461| / \sqrt{39.48192757}$

$= |-10.21165932| = 10.21165932$

(vi) Test of difference in mean of yolk ratio (x^)

$t = |54.85604 - 46.48875| / (\sqrt{(0.542491\ )^2/48) + (3.1\ 81622)^2/48))}$

$= |8.36729| / \sqrt{0.217021146}$

$= 17.961135$

Testing for the significance of Z value obtained at a = 0.1% significant table of t distribution with degree of freedom (v) assumed infinitely large we have t tabulated = 3.29.
Only variable x_3, x_4, x_5, and x_6 are significant at the preside significant level of a = 0.1%; since for each variable the $t_{calculated} > t_{tabulated}$ at v, a.

4.6 Test Of Equality Of Variance - Covariance Matrix

The null hypothesis H_0: $\hat{\Sigma}_1 = \hat{\Sigma}_2 = \hat{\Sigma}_3$ is tested against alternative hypothesis test H_1: $\hat{\Sigma}_1 \neq \hat{\Sigma}_2$.

The test statistic for the test is denoted by λ and is called the likelihood ratio test statistic for variance–covariance matrix.

The test statistic λ is given by $-2 \ln \lambda = (n_1 + n_2) \ln |\Sigma| - n_1 . |\Sigma_2|$ and it has chi-square distribution with ½ P(P + 1) degrees of freedom if the null hypothesis H_0 is true.

The estimate of λ is obtained as follows:
$-2 \ln \lambda = (48 + 48) \ln (1.63 \times 10^6 - 48 \text{ in } (2.5004 \times 101^{-6}) - 48 \ln (5.2403 \times 10^{-7}) - 2 \ln \lambda = 33.8520531$

$X = {}_e{}^{-16.92602655} = 4.45779 \times 10^{-8}$

The asymptotic null distribution of $-2 \ln\lambda$ is chi-square on ½ 6(6 + 1) = 21 degrees of freedom.
The tabulated value of chi-square distribution at a = 0.05 significance level is 8.03.
We thus conclude that the value of λ is not significant at a = 0.05 and we accept Ho that the dispersion matrices are the same and hence can be described by the pooled estimate ($\hat{\Sigma}$).

4.7 Results

The results obtained from the hypothesis testing reveals that the two groups under study can be distinguished from each other effectively using discriminant (discriminate function analysis) since the test of difference in mean shows a great significance for yolk dry weight (x_3), cholesterol/egg weight (x_5) yolk ratio (x_6) and slight significance for yolk cholesterol weight. The two groups can thus be distinguished based on these significant variables of interest.

Also, we observe that the dispersion matrices are the same such that $\hat{\Sigma}_1 = \hat{\Sigma}_2 = \hat{\Sigma}_3$ leads to a linear discriminant model given by:

llocate x to π_1 if $\hat{L}(x) > k$ and to π_2 if $\hat{L}(x) \leq k$

Where $\hat{L}(x) = (\mu^1 - \mu^{II})\ \hat{\Sigma}^{-1}[x - {}^1/_2\ (\mu^1 - \mu^{II})]$

For simplicity k is usually 1.

Using a more simplified expression for the discriminant model we have
$\hat{L}(x) = In[f_1(x)/f_2(x)] = L'(x) - \frac{1}{2} L'(\mu^1 - \mu^{II})$

Where $L = \hat{\Sigma}^{-1}(\mu^1 - \mu^{II})$

Evaluating $\hat{L}$ as the estimate of L we have
$\hat{L} = \hat{\Sigma}^{-1}(\overline{X}_1 - \overline{X}_2)$

$$\hat{L} = \hat{\Sigma}^{-1} \begin{pmatrix} 59.0997167 \\ 14.295625 \\ \\ 8.286875 \\ 10.99479 \\ 131.5671 \\ 54.85604 \end{pmatrix} \longleftrightarrow \begin{pmatrix} 58.35166667 \\ 14.7325 \\ \\ 6.89125 \\ 13.72083 \\ 195.7317 \\ 46.48875 \end{pmatrix}$$

While $\hat{\Sigma}^{-1}$ is the inverse of the pooled covariance matrix in the appendix.

$$\hat{L} = \begin{pmatrix} 800.2336216 \\ 986.1626 \\ -201.8453551 \\ -489.188092 \\ -1154.724668 \\ -3.340733547 \end{pmatrix}$$

Substituting the various estimates obtained from the data analysis previously we obtain a practical discriminant function model given by:

In $[f_1(x)/f_2(x)]$

$$= \begin{pmatrix} 800.2336216 \\ 986.1626 \\ -201.8453551 \\ -489.188092 \\ -1154.724668 \\ -3.340733547 \end{pmatrix} \begin{pmatrix} x \end{pmatrix} \begin{pmatrix} -\frac{1}{2} \end{pmatrix} \begin{pmatrix} 800.2336216 \\ 986.1626 \\ -201.8453551 \\ -489.188092 \\ -1154.724668 \\ -3.340733547 \end{pmatrix} \begin{pmatrix} \mu^1 - \mu^{II} \end{pmatrix}$$

Where $\hat{L}(X) = In\ [f_1(x)/f_2(x)]$

4.8 Reliability Of The Model

The reliability of the model is a question of how accurate the discriminant function model and allocation rule correctly allocates the groups under study as well as future groups.

$$\text{In}\,[f_1(x)/f_2(x)] = \begin{pmatrix} 800.2336216 \\ 986.1626 \\ -201.8453551 \\ -489.188092 \\ -1154.724668 \\ -3.340733547 \end{pmatrix} \begin{pmatrix} \\ x \\ \\ \\ \\ \end{pmatrix} \begin{pmatrix} \\ -½ \\ \\ \\ \\ \end{pmatrix} \begin{pmatrix} 8002336216 \\ 986.1626 \\ -201.8453551 \\ -489.188092 \\ -1154.724668 \\ -3.340733547 \end{pmatrix} \begin{pmatrix} \\ \mu^1 - \mu^{II} \\ \\ \\ \\ \end{pmatrix}$$

For organic copper salt, result that

$\overline{X}_1$ = [59.09979167, 14.295625, 8.286875, 10.99479, 131.5671, 54.85604]

Replacing x = $\overline{X}_1$ in the model we have In$[f_1(x)/f_2(x)]$ = -97766.77 + 135408.8623 = 37642.09233.

Also, for inorganic copper salt, we recall that

$\overline{X}_2$ = [58.35167, 14.7325, 6.89125, 13.72083. 195.7317, 46.48875]

Such that replacing $\overline{X}_2$ in the model we obtain

In$[f_1(x)/f_2(x)$ = -173050.9521 + 13508.8623 = -37642.08985

Since the value of In$[f_1(x)/f_2(x)]$ is positive for the organic copper salt group and negative for the inorganic copper salt group, then we conclude that the model is reliable and we perfectly classify future measurement to any of the two groups understudied.

Furthermore, for clarity of the reliability of the mode, each values in the data are substituted row-wise and the population to which they can be allowed are determined based on the value of In$[f_1(x)/f_2(x)]$ if it is negative or positive.

Example is in the row 20 of the organic copper salt data substituting x_1 = 61.06, x_2 = 14.68, x_3 = 8.37, x_4 = 11.6. x_5 = 161.37. x_6 = 55.38 yielded + 486.0372 which is a positive value showing that the information in row 210 of the organic copper salt data is correctly allocated.

Each row is considered and the percentage of individual observation not correctly allocated was determined.

Only about 12 individual observations are incorrectly allocated out of 48 observations in the organic copper salt data; while only about 1 out of 48 observations was incorrectly allocated in the inorganic copper salt data.

This implies that the % of the cost of misallocating an individual that suppose to be in the inorganic copper salt data to organic copper salt is about 25%.

The model still reliably allocate 75% of the data for the organic copper salt.

Also, the % of the cost of misallocating an individual that suppose to be in the organic copper salt data to the inorganic copper salt data section is 2.083%.

These suppose that the model still accurately and reliably predicts about 97.917% of the data for the organic copper salt.

With the discriminant model obtained (L(x)) being able to predict about 75% and 97.91% of organic and inorganic copper salt information; then we conclude that the model is accurate and reliable.

5. RECOMMENDATION AND CONCLUSION

5.1 Recommendation

Discriminant analysis and classification procedure is a reliable statistical tool which can be used in separating groups of objects with common variables of interests. As it is used in this project study to separate eggs produced into those that belong to organic group and inorganic group; it can be used in various fields of endeavour for similar effective result. Proper analysis of eggs produced by poultry operators so as to enhance good classification of eggs based on those with high cholesterol content and those with low cholesterol content. Based on the fact that high cholesterol content usually cause hear disease then I recommend eggs with low cholesterol content for adults prone to heart problems. The allocation rule used in any analysis should be observed for its weaknesses before been accepted for use. Transformation of data to normality is necessary in any discrimination and classification analysis since it makes the use of linear discriminant function simple.

5.2 Conclusion

In fact for several reasons, linear discriminant functions have been the focus of a considerable amount of research effort. First, it is obvious that linear discriminant functions are analytically simple. Secondly, they represent the simplest case (of more generalized discriminant functions) which has very wide applicability. Since most real life data's can be normalized by some statistical techniques and with the importance of central limit theorem on large data's, the application of linear discriminant functions becomes so wide and globally accepted only because the linear discriminant function is highly accurate for normally distributed data's.

Furthermore, we conclude that for every multivariate normally distributed data's the linear discriminant function is assumed when the test of variances shows equality and resolve to the use of quadratic discriminant function model when the dispersion matrices are different from each other. The livelihood test statistic for dispersion matrices is highly reliable for large sample size but changes to modification due to Box's M criterion when the sample size under study is small ($n < 30$). Discriminant analysis approach is a more positive tool for separation of groups even when the test of mean vectors between groups shows equality. This suggest that the test of equality of mean vectors is rather than uninteresting hypothesis that is in capable of revealing some hidden features required for separating in certain situations. Thus, in general, in statistics the right technical approach that can accurately reveal the hidden differences in the different groups is known as discriminant analysis.

The various allocation rule for new sets of data in any discriminant analysis and classification procedure have their advantages and drawbacks. Indeed, most allocation rule becomes approximately reliable when the sample size is convincingly large so that we then conclude which approach led to the lowest proportion of misclassification. In most cases the principle of cross valuation is a better method for assessing allocation rule than the re-substitution method.

The principle of cross validation reveals that the cholesterol content of eggs classified under the organic copper salt data are lower compared to that of inorganic copper salt data. The cholesterol content which is majorly determined by the variable x_4 and x_5 shows a classification rule given by allocate x to organic copper salt data section if $x_5 < 160$mg/egg and to inorganic copper salt data section if otherwise. Hence, we are able to separate eggs with high cholesterol content from those with low cholesterol content. This will successively reduce to the barest minimum heart problems caused by consumption of high cholesterol.

REFERENCES

1. Bakalli et al and Pesti 1996, An introduction to multivariate statistics. New York.
2. Fishers (1936) and conjufuca et al. (Scientific Researchers) Graphics Press, Cheshire Connection.
3. Harries R.J. (1985). A Princes on Multivariate Statistics 2nd Edition Academic Press Orlando.
4. Hotelling, H. (1933). Analysis of complex Statistical Variables into Principal Components. Journal of Educational Psychology.
5. Hnads D.J. (1981). Multivariate Statistics. Institute of Psychiatry London University.
6. Jackson J.E. (1991). A User's Guide to Principal Components Willey, New York.
7. Schultz, B. (1983). On Leven's and other Statistics of Variation. Evolutionary Theory, New York.
8. Seber G.A.F. (1984). Muitivariate Observations Willey, New York.
9. Srivastave M.S. and Carter E.M. (1983). An introduction to Applied Multivariate Statistics North-Holland, New York.
10. Tufte, E.R. (1983). The Visual Display of Quantitative information.