Comparison of Three Criteria for Discriminant Analysis Procedure

Nwosu, Felix D., Onuoha, Desmond O. and Eke Charles N.
Department of Mathematics and Statistics
Federal Polytechnic Nekede, Owerri, Imo State.
E-mail: desonuoha@yahoo.com

Abstract

This paper presents a fisher's criterion, Welch's criterion, and Bayes criterion for performing a discriminant analysis. These criteria estimates a linear discriminant analysis on two groups (or regions) of contrived observations. The discriminant functions and classification rules for these criteria are also discussed. A linear discriminant analysis is performed in order to determine the best criteria among Fisher's criterion, Welch's criterion and Bayes criterion by comparing their apparent error rate (APER). Any of these criteria with the least error rate is assumed to be the best criterion. After comparing their apparent error rate (APER), we observed that, the three criteria have the same confusion matrix and the same apparent error rate. Therefore we conclude that none of the three criteria is better than each other.

Kay Words: Fisher's criterion, Welch's criterion, Bayes criterion and Apparent Error rate

---

1. Introduction:

Discriminant Analysis is concerned with the problem of classification. This problem of classification arises when an investigator makes a number of measurements on an individual and wishes to classify the individual into one of several categories or population groups on the basis of these measurements. This implies that the basic problem of discriminant analysis is to assign an observation X, of more distinct groups on the basis of the value of the observation. In some problems, fairly complete information is available about the distribution of X in the two groups. In this case we may use this information and treat the problem as if the distributions are known. In most cases, however information about the distribution of X comes from a relatively small sample from the groups and therefore, slightly different procedures are used.

The Objectives of Discriminant Analysis includes: To classify cases into groups using a discriminant prediction equation; to test theory by observing whether cases are classified as predicted; to investigate differences between or among groups; to determine the most parsimonious way to distinguish among groups; to determine the percent of variance in the dependent variable explained by the independents; to assess the relative importance of the independent variables in classifying the dependent variable and to discard variables which has little relevance in relation to group distinctions.

In this study, we wish to determine the best criterion among the three criteria namely; Fisher's criterion, Welch's criterion and Bayes criterion for good discriminant functions, by comparing their apparent error rate (APER) while the significance is for detecting the variables that allow the researcher to discriminate between different groups and for classifying cases into different groups with a better than chance accuracy.

2. Related Literature:

Anderson[1] viewed the problem of classification as a problem of "statistical decision functions". We have a number of hypotheses which proposes is that the distribution of the observation is a given one. If only two populations are admitted, we have an elementary problem of testing one hypothesis of a specified distribution against another.

Lachenbruch (1975) viewed the problem of discriminant analysis as that of assigning an unknown observation to a group with a low error rate. The function or functions used for the assignment may be identical to those used in the multivariate analysis of variance.

Johnson and Wichern [23] defined discriminant analysis and classification as multivariate techniques concerned with separately distinct set of observations (or objects) and with allocating new observation (or object) to previously defined groups. They defined two goals namely: **Goal 1:** To describe either graphically (in at most three dimensions) or algebraically the differential features of objects (or observations) from several known collections (or populations) and **Goal 2:** To sort observations (or objects) into two or more labeled classes. The emphasis is on deriving a rule that can be used to optimally assign a new observation to the labeled classes. They used the term discrimination to refer to Goal 1 and used the term classification or allocation to refer to Goal 2. A more descriptive term for goal 1 is separation.

They also explained that a function that separates may sometimes serve as an allocator or classificatory and conversely an allocation rule may suggest a discriminator procedure. Also that goal 1 and 2 frequently overlap and the distinction between separation and allocation becomes blurred.

According to [2]; Discriminant function analysis is used to determine which variables discriminate between two or more naturally occurring groups.

Costanza W.J. and Afifi A.A. (1979) computationally stated that discriminant function analysis is very similar to analysis of variance (ANOVA).

Theoretical Basis by Lachenbruch P.A. [6] elaborated that the basic problem in discriminant analysis is to assign an unknown subject to one of two or more groups on the basis of a multivariate observation. It is important to consider the costs of assignment, the a priori probabilities of belonging to one of the groups and the number of groups involved. The allocation rule is selected to optimize some function of the costs of making an error and the a priori probabilities of belonging to one of the groups. Then the problem of minimizing the cost of assignment is to minimize the following equation

$$\text{Min} \sum \sum P\,(D_j/\prod_i)\, P_i\, C_{ji}$$

3.0 The Criterion:

3.1 Fishers Criterion:

Fisher (1936) suggested using a linear combination of the observations and choosing the coefficients so that the ratio of the differences of the means of the linear combination in the two groups to its variance is maximized. For classifying observation into one of two population groups, fisher considered the linear discriminant function

$y=\lambda^1 X$. Let the mean of y in population I ($\prod_1$) be $\lambda^1\mu_1$, and the mean of y in $\prod_2$ be $\lambda^1\mu_2$, its variance is $\lambda^1\sum\lambda$ in either population where $\sum = \sum_1 = \sum_2$. Then he chooses λ to

$$\text{Maximize } \Phi = \frac{\left(\lambda^1\mu_1 - \lambda^1\mu_2\right)^2}{\lambda^1\Sigma\lambda} \qquad (3.1.1)$$

Differentiating (3.1.1) with respect to λ, we have

$$\frac{d\Phi}{d\lambda} = \frac{2(\lambda^1\mu_1 - \lambda^1\mu_2)(\mu_1 - \mu_2)\lambda^1\Sigma\lambda - 2\lambda\Sigma(\lambda^1\mu_1 - \lambda^1\mu_2)^2}{\left(\lambda^1\Sigma\lambda\right)^2} \qquad (3.1.2)$$

Equating (3.1.2) to zero, we have

$$2(\lambda^1\mu_1 - \lambda^1\mu_2)(\mu_1 - \mu_2)\ \lambda^1\Sigma\lambda = 2\lambda\Sigma\ (\lambda^1\mu_1 - \lambda^1\mu_2)^2$$

$$\mu_1 - \mu = \frac{\Sigma\lambda(\lambda^1\mu_1 - \lambda^1\mu_2)}{\lambda^1\Sigma\lambda} \qquad (3.1.3)$$

Since λ is used only to separate the populations, we may multiply λ by any constant we desire. Thus λ is proportional to $\Sigma^{-1}(\mu_1 - \mu_2)$. The assignment procedure is to assign an individual to $\prod_1$, If $Y = (\mu_1 - \mu_2)^1\ \Sigma^{-1}X$ is closer to $\overline{Y_1} = (\mu_1 - \mu_2)^1\ \Sigma^{-1}\mu_1$ than to $\overline{Y_2} = (\mu_1 - \mu_2)^1\Sigma^{-1}\mu_2$ and an individual is assigned to Π_2 if $Y = (\mu_1 - \mu_2)^1$

Σ^{-1} X is closer to $\overline{Y_2}$ = $(\mu_1 - \mu_2)^1 \Sigma^{-1} \mu_1$ than to $\overline{Y_1}$. Then midpoint of the interval between $\overline{Y_1}$ and $\overline{Y_2}$ is

$$\frac{\overline{y_1} + \overline{y_2}}{2} = \tfrac{1}{2} (\mu_1 + \mu_2)^1 \Sigma^{-1} (\mu_1 - \mu_2).$$

This is used as the cut off point for the assignment. The difference between $\overline{Y_1}$ and $\overline{Y_2}$ is

$$\overline{Y_1} - \overline{Y_2} = (\mu_1 - \mu_2)^1 \sum^{-1} \mu_1 (\mu_1 - \mu_2)^1 \sum^{-1} \mu_2 = (\mu_1 - \mu_2)^1 \sum^{-1} (\mu_1 - \mu_2) = \delta^2$$

δ^2 is called the Mahalanobi's (squared) distance for known parameters. If the parameters are not known, it is the usual practice to estimate them by $\overline{X_1}$, $\overline{X_2}$ and S where $\overline{X_1}$ is the mean of a sample from $\prod_1$, $\overline{X_2}$ is the mean of a sample from $\prod_2$ and S is the pooled sample variance-covariance matrix from the two groups. The assignment procedure is to assign an individual to $\prod_1$ if $Y = (\overline{X_1} - \overline{X_2})^1 S^{-1} X$ is closer to $\overline{Y_1} = (\overline{X_1} - \overline{X_2})^1 S^{-1} \overline{X_1}$ than to $\overline{Y_2} = (\overline{X_1} - \overline{X_2})^1 S^{-1} \overline{X_2}$ while an individual is assigned to $\prod_2$ if $Y = (\overline{X_1} - \overline{X_2})^1 S^{-1} X$ is closer to $\overline{Y_2} = (\overline{X_1} - \overline{X_2})^1 S^{-1} \overline{X_2}$ than to $\overline{Y_1} = (\overline{X_1} - \overline{X_2})^1 S^{-1} \overline{X_1}$

The midpoint of the interval between $\overline{Y_1}$ and $\overline{Y_2}$ is $\frac{\overline{y_1} + \overline{y_2}}{2} = \tfrac{1}{2} (\overline{X_1} + \overline{X_2})^1 S^{-1} (\overline{X_1} - \overline{X_2})$.

Y is closer to $\overline{Y_1}$ if $|Y - \overline{Y_1}| < |Y - \overline{Y_2}|$ which occurs if $Y > \tfrac{1}{2} (\overline{Y_1} + \overline{Y_2})$ since $\overline{Y_1} > \overline{Y_2}$, The difference between $\overline{Y_1}$ and $\overline{Y_2}$ is $\overline{Y_1} - \overline{Y_2} = (\overline{X_1} - \overline{X_2})^1 S^{-1} \overline{X_1} - (\overline{X_1} - \overline{X_2})^1 S^{-1} \overline{X_2}$

$= (\overline{X_1} - \overline{X_2})^1 S^{-1} (\overline{X_1} - \overline{X_2}) = D^2$ which is called the Mahalanobis (squared) distance for unknown parameters.

The distribution of D^2 can be used to test if there are significant differences between the two groups (or Regions). We consider the two independent random samples (X_{ij}, j = 1, 2, . . . n_1) and (X_{2j}, j = 1, 2, . . . n_2) from $N_k(\mu_1, \sum)$ and $N_k(\mu_2, \sum)$ respectively. We test the hypothesis that both samples came from the same normal distribution, that is,

$$H0: \mu_1 = \mu_2 \text{ versus } H1: \mu_1 \neq \mu_2.$$

Let $A_1 = \sum_{j=1}^{n_1} (x_{1j} - \bar{x}_1)(x_{1j} - \bar{x}_1)^1$; $A_2 = \sum_{j=1}^{n_2} (x_{2j} - \bar{x}_2)(x_{2j} - \bar{x}_2)^1$

The pooled estimator S of $\sum$ is $S = \frac{A_1 + A_2}{n_1 + n_2 - 2} = \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2}{n_1 + n_2 - 2}$ and is unbiased for $\sum$

It is the property of the Wishart Distribution that if $X_{ij} \sim$ iid$N_k(N, \sum)$ $1 < j < n$ then

$A = \sum_{j=1}^{n} (x_{ij} - \bar{x}_1)(x_{ij} - \bar{x}_1)^1 \sim W_k (\sum, n\text{-}1)$; therefore $(n_1 + n_2 - 2)S \sim W_k(\sum, (n_1 + n_2 - 2)$ and is independent of $(\overline{X_1} - \overline{X_2})$ which is $N_k\left(0, \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\Sigma\right)$ when the null hypothesis is true.

$\frac{\overline{x_1} + \overline{x_2}}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim N_k(0, \Sigma)$ and is Independent of S. Therefore, $T^2 = V X^1 D^{-1} X$, V > k is the Hotelling's T^2 based on V degrees of freedoms where X and D are independent.

Here we have $X = \overline{X_1} - \overline{X_2}$ and $D = (n_1 + n_2 - 2)S$;

$$T^2 = \left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{-1} (\overline{x_1} - \overline{x_2})^1 ((n_1 + n_2 - 2)S)^{-1} (\overline{x_1} - \overline{x_2})(n_1 + n_2 - 2)$$

$$= \frac{n_1 n_2}{n_1 + n_2} (\overline{x_1} - \overline{x_2})^1 S^{-1} (\overline{x_1} - \overline{x_2}) \qquad (3.1.4)$$

If $X_2 \sim N_k(0, \sum)$ and $D \sim W_k(\sum, V)$, D, X is independent, then

$$T^2 \sim \frac{kv}{n - k + 1} F_{k, n-k+1}$$

Therefore,

$$T^2 = \frac{n_1 n_2}{n_1 + n_2}\left(\overline{x_1} - \overline{x_2}\right)^1 S^{-1}\left(\overline{x_1} - \overline{x_2}\right)$$

$$\sim \frac{kv}{n-k+1} F_k, n-k+1$$

$$F = \frac{n_1 n_2 (n_1 + n_2 - k - 1)}{n_1 + n_2 (n_1 + n_2 - 2)k}\left(\overline{x_1} - \overline{x_2}\right)^1 S^{-1}\left(\overline{x_1} - \overline{x_2}\right)$$

$$F = \frac{n_1 n_2 (n_1 + n_2 - k - 1)}{n_1 + n_2 (n_1 + n_2 - 2)k} D^2 \qquad (3.1.5)$$

The variable $F = \frac{n_1 n_2 (n_1 + n_2 - k - 1)}{n_1 + n_2 (n_1 + n_2 - 2)k} D^2$

where n_1 and n_2 are the sample sizes in $\prod_1$ and $\prod_2$ respectively and K is the number of variables, has an F-distribution with F and $n_1 + n_2 - k - 1$ degrees of freedom. The use of $\frac{\overline{y_1} + \overline{y_2}}{2}$ as a cut off point can be improved upon if the apriori probabilities of $\prod_1$ and $\prod_2$ are not equal.

3.2 Welch's Criterion

An alternative way to determine the discriminant function is due to Welch (1939). Let the density functions of $\prod_1$ and $\prod_2$ be denoted by $F_1(X)$ and $F_2(X)$ respectively. Let q_1 be the proportion of $\prod_1$ in the population and $q_2 = (1 - q_1)$ be the proportion of $\prod_2$ in the population. Suppose we assign X to $\prod_1$ if X is in some region R_1 and to $\prod_2$ if X is in some region R_2. We assume that R_1 and R_2 are mutually exclusive and their union includes the entire space R. The total probability of misclassification is,

$$T(R,F) = q_1 \int_R f_1(x)dx + q_2 \int_R f_2(x)dx$$

$$= q_1 (1 - \int_R f_1(x)dx) + q_2 \int_R f_2(x)dx$$

$$= q_1 + \int_R (q_2 f_2(x) - q_1 f_1(x))dx \qquad (3.2.1)$$

This quantity is minimized if R_1 is chosen such that $q_2 f_2(x) = q_1 f_1(x) < O$ for all points in R_1

Thus the classification rule is:

Assign X to $\prod_1$ if $\frac{f_1(x)}{f_2(x)} > \frac{q_2}{q_1}$ (3.2.2)

And to $\prod_2$ if otherwise; it is pertinent to note that this rule minimizes the total probability of misclassification.

An important special case is when $\prod_1$ and $\prod_2$ are multivariate normal with means μ_1 and μ_2 and common covariance matrix $\sum$. The density in population $\prod_1$ is

$$f_i(x) = \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} \exp\{-1/2(x-\mu_i)^1 \Sigma^1 (x-\mu_i)\} \qquad (3.2.3)$$

The ratio of the densities is

$$\frac{f_1(x)}{f_2(x)} = \frac{\exp\{-1/2(x-\mu_1)^1 \Sigma^1 (x-\mu_1)\}}{\exp\{-1/2(x-\mu_2)^1 \Sigma^1 (x-\mu_2)\}}$$

$$= \exp[-1/2\{(x-\mu_1)^1 \Sigma^1 (x-\mu_1) - (x-\mu_2)^1 \Sigma^1 (x-\mu_2)\}]$$

$$= \exp\left\{1/2\left[-x^1\Sigma^{-1}\mu_1 - \mu_1\Sigma^{-1}x^1 + \mu_1\Sigma^{-1}\mu_1 + x^1\Sigma^{-1}\mu_2 + \mu_2\Sigma^{-1}x^1 - \mu_2\Sigma^{-1}\mu_2\right]\right\}$$

$$= \exp\left\{\left[X - 1/2(\mu_1 + \mu_2)\right]^1 \Sigma^{-1}(\mu_1 - \mu_2)\right\} \qquad (3.2.4)$$

The optimal rule is to assign the unit X to $\prod_1$ if

$$= D_T(X) = \left[X - 1/2(\mu_1 + \mu_2)\right]^1 \Sigma^{-1}(\mu_1 - \mu_2) > in\frac{q_2}{q_1} \qquad (3.2.5)$$

The quantity on the left of equation 3.2.5 is called the true discriminant function $D_T(X)$. Its sample analogue is

$$D_T(X) = \left[X - 1/2\left(\overline{X}_1 + \overline{X}_2\right)\right]^1 S^{-1}\left(\overline{X}_1 - \overline{X}_2\right) \qquad (3.2.6)$$

The coefficient of X is seen to be identical with Fishers result for the linear discriminant function. The function $D_T(X)$ is a linear transformation of X and knowing its distribution will make it possible to calculate the error rates that will occur if $D_T(x)$ is used to assign observation to $\prod_1$ and $\prod_2$. Since X is multivariate normal, $D_T(x)$ being a linear combination of X is normal. The means of $D_T(x)$ if X comes form $\prod_1$ is

$$E\left(D_T(x)/\Pi_1\right) = \left[\mu_1 - 1/2(\mu_1+\mu_2)\right]^1 \Sigma^{-1}(\mu_1-\mu_2) \quad \frac{q_1 f_1(x)}{q_1 f_1(x)+q_2 f_2(x)} > \frac{q_2 f_2(x)}{q_1 f_1(x)+q_2 f_2(x)} \quad (3.3.1)$$

$$= \left[\mu_2 - 1/2\,\mu_1 - 1/2\,\mu_2\right]^1 \Sigma^{-1}(\mu_1-\mu_2)$$
$$= -1/2(\mu_1-\mu_2)^1 \Sigma^{-1}(\mu_1-\mu_2)$$
$$= 1/2\,\delta^2$$

Where $\delta^2 = (\mu_1 - \mu_2)^1 \sum^{-1} (\mu_1 - \mu_2)$
In $\prod_2$, the mean of $D_T(x)$ is

$$E\left(D_T(x)/\Pi_1\right) = \left[\mu_1 - 1/2(\mu_1+\mu_2)\right]^1 \Sigma^{-1}(\mu_1-\mu_2)$$

$$= \left[\mu_2 - 1/2\,\mu_1 - 1/2\,\mu_2\right]^1 \Sigma^{-1}(\mu_1-\mu_2)$$
$$= -1/2(\mu_1-\mu_2)^1 \Sigma^{-1}(\mu_1-\mu_2)$$
$$= 1/2\,\delta^2$$

In either population the variance is

$$ED_T(x) - D_T(\mu_i)^2 = E(\mu_1-\mu_2)^1 \Sigma^{-1}(x-\mu_i)(x-\mu_i)^1 \Sigma^{-1}(\mu_1-\mu_2)$$

$$= (\mu_1-\mu_2)^1 \Sigma^{-1} E(x-\mu_i)(x-\mu_i)^1 \Sigma^{-1}(\mu_1-\mu_2)$$
$$= (\mu_1-\mu_2)^1 \Sigma^{-1}(\mu_1-\mu_2) = \delta^2$$

The quantity δ^2 is the population Mahalanobis (squared) distance.

3.3 Bayes Criterion

A Bayesian criterion for classification is one that assigns an observation to a population with the greatest posterior probability. A Bayesian criterion for classification is to place the observation in $\prod_1$ if p ($\prod_1$/x) > p ($\prod_2$/x).

By Bayes theorem

$$p\left(\Pi_i/x\right) = p(\Pi : x)/p(x)$$
$$= \frac{q_1 f_i(x)}{q_1 f_1(x) + q_2 f_2(x)};$$

Hence the observation X is assigned to $\prod_1$ if

The above rule reduces to assigning the observation to $\prod_1$ if $\frac{f_1(x)}{f_2(x)} > \frac{q_2}{q_1}$ (3.3.2)

and to $\prod_2$ otherwise.

Where

$$\frac{f_1(x)}{f_2(x)} = \exp\left\{\left[X - 1/2(\mu_1+\mu_2)\right]^1 \Sigma^{-1}(\mu_1-\mu_2)\right\}$$

q_1 = the proportion of $\prod_1$ in the population.
q_2= (1-q_1) = the proportion of $\prod_2$ in the population.

3.4 Probabilities of Misclassification

In constructing a procedure of classification, it is desired to minimize the probability of misclassification or more specifically, it is desired to minimize on the average the bad effects of misclassification. Suppose we have an observation from either population $\prod_1$ or population $\prod_2$ the classification of the observation depends on the vector of measurements.

X^1= (X_1, X_2, . . . , X_k)^1 on the observation. We set up a rule that if an observation is characterized by certain sets of values of X_1, X_2,. . .,X_k, we classify it as from $\prod_1$, if it has other values, we classify it as from $\prod_2$. We think of an observation as a point in a K-dimensional space. We divide the space into two regions or groups. If the observation falls in R_1, we classify it as coming from population $\prod_1$, and if it falls in R_2, we classify it as coming from population $\prod_2$.

In following a given classification procedure, the statistician can make two kinds of errors in classification. If the observation is actually from $\prod_1$, the statistician or researcher can classify it as coming from $\prod_2$; or if it is from $\prod_2$, the statistician may classify it as from $\prod_1$. We need to know the relative undesirability of these two kinds of misclassification.

Let the probability that an observation comes from population $\prod_1$ be q_1 and from population $\prod_2$ be q_2. Let the density function

of population $\prod_1$ be $f_1(x)$ and that of population $\prod_2$ be $f_2(x)$. Let the regions of classification from $\prod_1$ be R_1 and from $\prod_2$ be R_2. Then the probability of correctly classifying an observation that is actually drawn from $\prod_1$ is $\int_{R_1} f(x)dx$ where $dx = dx_1, dx_2, \ldots, dx_k$ and the probability of misclassifying an observation from $\prod_1$ is $P_1 = \int_{R2} f(x)dx$

Similarly the probability of correctly classifying an observation from $\prod_2$ is $\int f_2(x)\,dx$ and the probability of misclassifying such an observation is $P_2 = \int_{R2} f(x)dx$; then the total probability of misclassification is

$$T(R;f) = q_1 \int_{R2} f_1(x)dx + q_2 \int_{R1} f_2(x)dx \quad (3.4.1)$$

Table 3.4.1: confusion matrix

Statisticians' decision

Population		Π_1	Π_2
	Π_1	Correct Classification	P_1
	Π_2	P_2	Correct Classification

Probabilities of misclassification can be computed for the discriminant function. Two cases have been considered.

(i) When the population parameter are know.

(ii) When the population parameter are not known but estimated from samples drawn from the two populations.

3.5 Apparent Error Rates (APER)

One of the objectives of evaluating a discriminant function is to determine its performance in the classification of future observations. When the (APER) is

$$T(R;f) = q_1 \int_{R2} f_1(x)dx + q_2 \int_{R1} f_2(x)dx$$

If $f_1(x)$ is multivariate normal with mean μ_1 and covariance $\sum$, we can easily calculate these rates. When the parameters are not known a number of error rates may be defined. The function T(R,F) defines the error rates (APER). The first argument is the presumed distribution of the observation that will be classified.

4.0 Data Analysis

Consider to carry out a linear discriminant analysis on two groups (or regions) of contrived observations.

A

X_1	6	7	9	8	8	10
2	7	5	10	8	9	9

B

X_1	11	15	22	17	12	13
X_2	13	16	20	16	11	14

4.1: Using Fishers Criterion
For A

$$A = \begin{pmatrix} \Sigma x_1^2 - N\mu_1^2 & \Sigma x_1 x_2 - N\mu_1\mu_2 \\ \Sigma x_2 x_1 - N\mu_2\mu_1 & \Sigma x_2^2 - N\mu_2^2 \end{pmatrix} \qquad = A = \begin{pmatrix} 10 & 9 \\ 9 & 16 \end{pmatrix}$$

For A

$\Sigma X_1 = 48, \Sigma X_2 = 48, \Sigma X_1^2 = 394, \Sigma X_2^2 = 400, \Sigma X_1 X_2 = 393, \overline{X_1} = \mu_1 = 8, \overline{X_2} = \mu_2 = 8, N = 6$

For B

$\Sigma X_1 = 90, \Sigma X_2 = 90, \Sigma X_1^2 = 1432, \Sigma X_2^2 = 1398, \Sigma X_1 X_2 = 1409, \overline{X_1} = \mu_1 = 15, \overline{X_2} = \mu_2 = 15, N = 6$

$$B = \begin{pmatrix} \Sigma x_1^2 - N\bar{x}_1^2 & \Sigma x_1 x_2 - N\bar{x}_1\bar{x}_2 \\ \Sigma x_2 x_1 - N\bar{x}_2\bar{x}_1 & \Sigma x_2^2 - N\bar{x}_2^2 \end{pmatrix} \qquad => \qquad B = \begin{pmatrix} 82 & 59 \\ 59 & 48 \end{pmatrix}$$

$$S = \frac{A+B}{n_1 + n_2 - 2} \qquad = \qquad \begin{pmatrix} 9.2 & 6.8 \\ 6.8 & 6.4 \end{pmatrix}$$

$$Y = \left(\overline{x_1} - \overline{x_2}\right)^1 S^{-1} X$$

$Y = 0.2212X_1 - 1.3293X_2$ which is the discriminant function.

$$\overline{Y_1} = \left(\overline{x_1} - \overline{x_2}\right)^1 S^{-1} \overline{X_1} \qquad => \qquad \overline{Y_1} = (0.2212 - 1.3293)\begin{pmatrix} 8 \\ 8 \end{pmatrix} = -8.8648$$

$$\overline{Y_2} = (0.2212 - 1.3293)\begin{pmatrix} 15 \\ 15 \end{pmatrix} = -16.6215$$

Cut off point $= \dfrac{\overline{Y_1} + \overline{Y_2}}{2}$ and this is also referred to as the mid point and it's equal to -12.74315.

Assignment procedure:

Assign observation with measurement X to $\prod_1$ if $Y > \dfrac{\overline{Y_1} + \overline{Y_2}}{2}$ and assign to $\prod_2$ if $Y \leq \dfrac{\overline{Y_1} + \overline{Y_2}}{2}$

Discriminant scores

$Y = 0.2212X_1 \quad -1.3293X_2$

A	-7.9779	-5.0981	-11.3022	-8.8648	-10.1941	-9.7517
B	-14.8477	-17.9508	-21.7196	-18.8356	-17.5084	-15.7346

For Group A
-7.9779 – (-12.74315) = 4.7653 > $\prod_1$
-5.0981 – (-12.74315) = 7.6451 > $\prod_1$
-11.3022 – (-12.74315)= 1.4410 > $\prod_1$

For Group B
-14.8477 – (-12.74315) = -2.1046 < $\prod_2$
-16.9508 – (-12.74315) = -5.2077 < $\prod_2$
-21.7196 – (-12.74315) = -6.0925 < $\prod_2$

-8.8648 – (-12.74315) = 2.5491 > $\prod_1$ -18.8356- (-12.74315) = -6.0925 < $\prod_2$
-10.1941 – (-12.74315)= 2.5491 > $\prod_1$ -17.5084 – (-12.74315) = 0.7753 > $\prod_1$
-9.7517 – (-12.74315) = 2.9915 > $\prod_1$ -15.7346 – (-12.74315) = -2.9915 < $\prod_2$

Tables 4.1.1. Confusion matrix
Statistician decision

	$\prod_1$	$\prod_2$
Population $\prod_1$	6	0
Population $\prod_2$	1	5

The probability of misclassification
P(2/1) = 0/6 = 0
P (1/2) = 1/6

Apparent error rate (APER)

Error rate $= \dfrac{n\left(\frac{2}{1}\right)+n\left(\frac{1}{2}\right)}{Total\Pi_1 + Total\Pi_2} = \dfrac{1}{12}$; hence the APER $= \dfrac{1}{12}$

4.2 : Using Welch's Criterion

The classification rule is:

Assign X to $\prod_1$ if $\dfrac{f_1(x)}{f_2(x)} > \dfrac{q_2}{q_1}$ and to $\prod_2$ if otherwise.

$$\frac{f_1(x)}{f_2(x)} = \exp\left\{\left[X - \tfrac{1}{2}(\mu_1+\mu_2)\right]^1 \Sigma^{-1}(\mu_1-\mu_2)\right\} > \frac{q_2}{q_1}$$

Taking the Lim of both sides; we have $\left\{\left[X - \tfrac{1}{2}(\mu_1+\mu_2)\right]^1 \Sigma^{-1}(\mu_1-\mu_2)\right\} > In\dfrac{q_2}{q_1}$ therefore;

$$D_T(x) = \left\{\left[X - \tfrac{1}{2}(\mu_1+\mu_2)\right]^1 \Sigma^{-1}(\mu_1-\mu_2)\right\} > In\frac{q_2}{q_1}$$

Where
$D_T(x)$ is called the true discriminant function and $q_1 = q_2$ since they have equal sample size.
The optimal rule is to assign the unit X to $\prod_1$ if

$$D_T(x) = \left\{\left[X - \tfrac{1}{2}(\mu_1+\mu_2)\right]^1 \Sigma^{-1}(\mu_1-\mu_2)\right\} > In\frac{q_2}{q_1}$$ and to $\prod_2$ if otherwise.

But $q_1 = \dfrac{n_1}{n}$ where $n = n_1 + n_2 = 6+6 = 12$ and $q_2 = \dfrac{n_2}{n}$ where $n = n_1 + n_2 = 6+6 = 12$, hence;

$$q_1 = q_2 = \frac{1}{2}$$

$$D_T(x) = \left\{\left[X - \tfrac{1}{2}\left(\overline{X_1}+\overline{X_2}\right)\right]^1 S^{-1}\left(\overline{X_1}-\overline{X_2}\right)\right\}$$ while the $S^{-1} = \begin{pmatrix} 0.5063 & -0.5379 \\ -0.5379 & 0.7278 \end{pmatrix}$

$$D_T(x) = \left\{\left[X - \tfrac{1}{2}\left(\overline{X_1}+\overline{X_2}\right)\right]^1 S^{-1}\left(\overline{X_1}-\overline{X_2}\right)\right\} > In\frac{0.5}{0.5} \Rightarrow In(1) > 0$$

A B

X_1	6	7	9	8	8	10
X_2	7	5	10	8	9	9

X_1	11	15	22	17	12	13
X_2	13	16	20	16	11	14

Table 4.2.1: Confusion Matrix
Statistician Decision

	$\prod_1$	$\prod_2$
Population $\prod_1$	6	0
Population $\prod_2$	1	5

The probability of misclassification
P(2/1) = 0/6 = 0
P (1/2) = 1/6

$$\textbf{Error rate} = \frac{n\left(\frac{2}{1}\right) + n\left(\frac{1}{2}\right)}{Total\Pi_1 + Total\Pi_2} = \frac{1}{12}$$

4.3: Using Bayes Criterion

The classification rule:

Observation X is assigned to $\prod_1$ if $\frac{f_1(x)}{f_2(x)} > \frac{q_2}{q_1}$ and to $\prod_2$ if otherwise.

$$\frac{f_1(x)}{f_2(x)} = \exp\left\{\left[X - \tfrac{1}{2}(\mu_1 + \mu_2)\right]' \Sigma^{-1}(\mu_1 - \mu_2)\right\} > \frac{q_2}{q_1}$$

Note that $q_1 = q_2 = \frac{1}{2}$ which means that $\frac{q_2}{q_1} = 1$

$$= \exp\left\{\left[X - \tfrac{1}{2}(\mu_1 + \mu_2)\right]' \Sigma^{-1}(\mu_1 - \mu_2)\right\} > \frac{q_2}{q_1}$$

$$S^{-1} = \Sigma^{-1} = \begin{pmatrix} 0.5063 & -0.5379 \\ -0.5379 & 0.7278 \end{pmatrix}$$

Therefore, observation X is assigned to $\prod_1$ if $\exp\left\{\left[X - \tfrac{1}{2}\left(\overline{X_1} + \overline{X_2}\right)\right]' S^{-1}\left(\overline{X_1} - \overline{X_2}\right)\right\} > 1$ and to $\prod_2$ if otherwise.

Table 4.3.1 Confusion Matrix
Statistician Decision

	$\prod_1$	$\prod_2$
Population $\prod_1$	6	0
Population $\prod_2$	1	5

The probability of misclassification
P(2/1) = 0/6 = 0
P (1/2) = 1/6

$$\textbf{Error rate} = \frac{n\left(\frac{2}{1}\right) + n\left(\frac{1}{2}\right)}{Total\Pi_1 + Total\Pi_2} = \frac{1}{12} = 0.0833$$

5.0 Summary, Conclusion and Recommendation

5.1: Summary

Discriminant Analysis and Classification is defined by Johnson and Wichern [23] as multivariate techniques concerned with separating district set of objects and with allocating new objects to previously defined groups.

In fisher's criterion, object X is assigned to population $\prod_1$ if $Y > \frac{\overline{Y_1} + \overline{Y_2}}{2}$ and to $\prod_2$ if otherwise; and in Welch's criterion, the optimal rule is to assign the unit X to $\prod_1$ if $D_T(x) =$ $\left\{\left[X - \frac{1}{2}(\mu_1 + \mu_2)\right]' \Sigma^{-1}(\mu_1 - \mu_2)\right\} > In\frac{q_2}{q_1}$ and to $\prod_2$ if otherwise while in Bayes theorem, the object X is assigned to $\prod_1$ if $\frac{f_1(x)}{f_2(x)} > \frac{q_2}{q_1}$ and to $\prod_2$ if otherwise.

5.2: Conclusion

In other to know the best criteria among fisher's criterion, Welch's criterion, and Bayes criterion, we carried out a linear discriminant analysis on two groups (or regions) of contrived object (or observations). After the analysis, we discovered that the three criteria (Fisher's criterion, Welch's criterion, and Bayes criterion) had equal error rate, that is, none of them is better than each other in linear discriminant analysis.

5.3: Recommendation

We recommend for further studies with enlarged sample size to ascertain if the conclusion can be validated.

References

[1]Anderson T.W (1973) "Asymptotic evaluation of the probabilities of misclassification by linear discriminant functions" In Discriminant analysis and applications, T.Cacoullos edition New York Academic press page 17-35.

[2]Bartlelt M.S. (1951) "An inverse matrix adjustment arising in discriminant analysis" Annals of Mathematical Statistics, 22 page 107-111.

[3]Costanza W.J. and Afifi A.A. (1979)"Comparison of stopping rules in forward stepwise discriminant analysis" Journal of American Statistical Association, 74, page 777-785.

[4]Lachenbruch P.A. (1968) "On the expected values of probabilities of misclassification in discriminant analysis, necessary size and a relation with the multiple correlation coefficient" Biometrics 24, page 823.

[5]Lachenbruch P.A. (1975) Discriminant Analysis. Hafner press New York.

[6] Lachenbruch P.A. and Mickey M.R. (1968) "Estimation of Error Rates in Discriminant Analysis" Technometrics, 10, page 1.

[7] Onyeagu S.I. and Adeboye O.S. (1996) "Some methods of Estimating the probability of misclassification in Discriminant Analysis" Journal of the mathematical Association of Nigeria ABACUS vol 24 no 2 page 104-112.

[8] Onyeagu Sidney I. (2003): "A first Course in Multivariate Statistical Analysis", A Profession in Nigeria Statistical Association, Mega concept Publishers, Awka, Page. 208-221.

[9] Smith C.A.B. (1947) "Some examples of discrimination" Annals of Eugenics, 18 page 272-283.