Iraqi Journal of Statistical Sciences

http://stats.uomosul.edu.iq

Fuzzy Discriminant Analysis with Application

Rafal Talal Saadi [1] Alla A. Hamoodat[2]

[1,2]Department of Statistics and Informatics, College of Computer Science and Mathematics University of Mosul, Mosul, Iraq

Article information

Article history:
Received :January 1,2025
Revised: March 30, 2025
Accepted :June 2,2025
Available online: December 1,2025

Keywords:
Discriminant Analysis,
Hoteling T2,
Fuzzy Logic,
Cutting Point

Correspondence:
Alla A. Hamoodat
allahamoodat@uomosul.edu.iq
rafaltalal02@gmail.com

Abstract

In the current research, two methods of data classification were used, namely; the Linear Discriminant analysis (LDA) and the Fuzzy Discriminant Analysis (FDA). The discriminant analysis is considered as a method in which a certain item is analyzed into one of the sets depending on statistically significant variables in the process of analysis. The item is classified within the suitable population through a set of variables. Sometimes, the data is non-linear or does not follow a normal distribution or is unstable, so the use of fuzzy logic is successful because it does not require the data to follow a normal distribution or to be stable and Through fuzzy theory, models are developed to solve problems that cannot be analyzed using pure mathematical methods. The research aims to compare the two analysis methods (LDA) and (FDA) in order to display data and distinguish between different categories with high accuracy, provide accurate classifications of observations, as well as diagnose variables of high importance in discriminatory data using confidence intervals for the Roy-Bose test and the t-test. Real data was used for two groups of kidney patients (infected, not infected) depending on a set of influencing factors and the extent to which each factor affects the persons or the new items and their distribution on one of the two sets, which in turn, leads to early diagnosis and avoiding the deterioration of the health state of the patient. Results were obtained after applying the statistical package (SPSS) and (R) statistical package, which indicated that the fuzzy discriminant analysis (FDA) was the best analysis as it yields the least classification error by depending on the mean standard error (MSE).

DOI:10.33899/iqjoss.v22i2.54083,

1.Introduction

The recent years witnessed a considerable increase in the number of the persons who suffer kidney failure and this disease has a serious complications that affect the heart, arteries, nerves, kidneys and eyes. This disease has always been the reason behind the rise in mortalities directly or indirectly and the reason behind that is the pancreas inability to secrete the quantity of insulin that is required to do the functions or secreting ineffective insulin that leads to disorders including the rise of sugar in the blood, which, in turns, results in damages in the cardiovascular and eyes. The classification of a new single observation of one of the groups in question is done through the discriminant analysis, which is one of the important statistical methods, as a set of variables are used between two sets or more through using the linear discrimination function and the fuzzy discrimination function. The discriminant analysis is considered one of the vital types of analysis for the statisticians especially in the medical

field that is characterized with having many variables, that it makes it difficult to the researcher to analyze and extract the results accurately only (Oladapo, et al., 2024).

The membership of each single observation is related to one of the sets under study in accordance with the discrimination analysis by obtaining the Cut-Point that separates the sets in question based on the weights and ratios that are obtained by finding a certain function of the explanatory variables that are predicted according to the linear discriminant analysis and the fuzzy discriminant analysis. The current research aims at determining the best discriminant function depending on the Mean Statistical Error, i.e. depending on the discrimination function that classifies a new item to one of the sets with least possible classification error.

2. Material and Methods

2.1 Linear Discriminant Function (LDF)

This function is defined as a mathematical model that can be formulated through the indicators of a sample with observations that were selected randomly from two different sets. This function enables us to test any single observation and determine its set this type of discrimination is considered one of the simplest types of discrimination that hypothesizes that the explanatory variables are normally distributed with multivariable and the matrices of variance and pool variance are equal and this means the acceptance of the null hypothesis when testing the hypothesis (Afifi and Clark, 1984):

$H_0: \Sigma_1 = \Sigma_2 = \ldots = \Sigma_k$

$H_1: \Sigma_1 \neq \Sigma_2 \neq \ldots \neq \Sigma_k$

As

Σ: the matrices of variance and pool variance.

k: the number of sets

Hypothesizing that there are two sets, we refer to the first set as (1) and the second set (2), the values of m observations of the random variables that can be depended upon in clasification, which are X_1 , X_2 ,…, X_m. So, the classification function becomes:

$$Z = \alpha_1 X_1 + \alpha_2 X_2 + \ldots + \alpha_m X_m \quad (1)$$

Where;

α: the parameters of the model that are used in the classificationn process.

m: the number of variables.

2.2 Testing the Significance of the Linear Discriminant Function

The linear discriminant function is tested in two ways:

First: Statistical Hotelling T^2

When one wants to test the discrimination between two sets and forming the discrimination function with significantly accceptable, i.e. there are significant differences between the averages of the sets, then we test the following hypothesis (**Elkhouli, 2024)**

$H_0: \Sigma_1 = \Sigma_2$

$H_1: \Sigma_1 \neq \Sigma_2$

In order to test the hypothesis, we use the statistics of (Hoteeling^2) as this indicator is used to deal with the statistical problem that is related to the decision making about the two samples that have multivariable normal distribution and have the same pool variance matrix and this indicator is a development of (t) test from one sample to multivariable locations with. The (t) test

is used when comparing the mean of two normal populations and the hypothesis of this test is (Qu & Pei ,2024) (Karzan, F. , et al ,2024):

$H_0 : \mu_1 - \mu_2 = 0$

$H_1 : \mu_1 - \mu_2 \neq 0$

The standard of the hypothesis testing is:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_P \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad (2)$$

$$S_p^2 = \frac{(n_1 - 1)S_{x_1}^2 + (n_2 - 1)\, S_{x_2}^2}{n_1 + n_2 - 2} \quad (3)$$

Where;

S_p^2 : the pooled variance.

$(\bar{X}_1 - \bar{X}_2)$: the difference between the averages of the two samples.

$S_{x_1}^2$ variance for the first sample.

$S_{x_2}^2$: variance for the second sample.

(n_1, n_2): the sizes of the first and the two samples respectively.

H_0 is rejected if Cal.t > tab.t = t(1-$\frac{\alpha}{2}, n_1 + n_2 - 2$) is at the significance of α.

Taking into consideration that the statistics of Hotelling T^2 is suitable with the statistics of (Mahalanobis D^2) and depends on it (Arnold, 1981) (Adebayo, et al., 2024).:

$$D^2 = [\, \bar{X}_1 - \bar{X}_2 \,]' S^{-1} \, [\bar{X}_1 - \bar{X}_2 \,] \quad (4)$$

$$T^2 = \frac{n_1\, n_2}{n_1 + n_2} D^2 \quad (5)$$

Due to the difficulty of extracting the table value as there are no relevant tables for this purpose the table value of (F) can be obtained directly, which was stipulated by (Rao) in 1952 as (F) test is related directly to the statistics of Hotteling T^2:

$$F = \frac{n_1 + n_2 - p - 1}{(n_1 + n_2 - 2)p} T^2 \quad (6)$$

Then F table value is obtained at a significance level of (α) and with a degree of freedom of tab.F=F(α,p,n_1,+n_2-m-1). So, if Cal.F >= tab.F, then the null hypothesis (H_0) is rejected and the alternative hypothesis is accepted at a level of significance of (α) and this denotes that the coefficients used in the classification can be relied upon in the classification.

Second: Testing the Significance of the Discriminant Function using the Variance Analysis

The focus is concentrated on the discriminant function and its validity for discrimination and this can be determined through the regression analysis and the hypothesis of this test is as follows (Al-Rawi, 1987):

$H_0: \alpha_1 = \alpha_2 = \ldots = \alpha_k$

$H_1: \alpha_1 \neq \alpha_2 \neq \ldots \neq \alpha_k$

Depending on the (D^2) value, we find that the Sum Squares Within sets is:

$$SS(between) = \left(\frac{n_1\, n_2}{(n_1 + n_{2)\,(n_{1+}\, n_2\, -2)}}\right)(D^2)^2 = \frac{T^2\, D^2}{n_{1+}\, n_2\, - 2} \quad (7)$$

Then we find the Sum Squares Within sets:

$SS(between) = D^2$

So, the average of the squares with the sets is:

$$MS(between) = \frac{SS(B)}{m}$$

$$MS(within) = \frac{D^2}{(n_1 + n_2) - (m + 1)}$$

The following table of variance shows that and from it the F value is determined as shown in table (1):

Table (1): The variance analysis of the discriminant analysis

S.O.V	$d.f$	$S.S$	$M.S$	$cal.F$
Between Sets	m	$SS(B)$	$MS(B)$	$\frac{MS(B)}{MS(W)}$
Within Sets Error	$n_1 + n_2 - (m + 1)$	$SS(W)$	$MS(W)$	
Total	$n_1 + n_2 - 1$	$SS(T)$		

From Table (1), the calculated value (Cal.F) is compared with the tabulated value (tab.F) at the significance level (α) and the degree of freedom tab.F=F(α,p,n_1,+n_2-m-1). It is noted that the value of (Cal. F) is greater than (tab. F), and this indicates that the function is important and valid for discrimination, and this means that there is at least a significant variable.

2.3 The Cut-Point

To make the discrimination function in equation (1) as a tool for classifying the individuals, we need a point that separates the two sets. The following equation is used to find the discrimination function after the significant variables are determined that should be entered into the function, as the cut point (separation) is determined using applying the following equation (Afifi and Clark, 1984):

$$CutPoint(CP) = \frac{\bar{Z}_1 + \bar{Z}_2}{2}$$

$$CP = \frac{\left[(\bar{\underline{X}}_1 - \bar{\underline{X}}_2)' \underline{S}^{-1} \bar{\underline{X}}_1\right] + \left[(\bar{\underline{X}}_1 - \bar{\underline{X}}_2)' \underline{S}^{-1} \bar{\underline{X}}_2\right]}{2}$$

$$CP = \frac{\left[(\bar{\underline{X}}_1 - \bar{\underline{X}}_2)' \underline{S}^{-1} (\bar{\underline{X}}_1 + \bar{\underline{X}}_2)\right]}{2} \quad (8)$$

To facilitate the classification, the cut point can be subtracted from the value that the discriminant function gives and therefore the cut point will be the zero, as follows:

$$\therefore W = Z - CP$$

$$\therefore W = X' \underline{S}^{-1} (\bar{\underline{X}}_1 - \bar{\underline{X}}_2) - \frac{1}{2} (\bar{\underline{X}}_1 + \bar{\underline{X}}_2)' \underline{S}^{-1} \left(\bar{\underline{X}}_1 - \bar{\underline{X}}_2\right) \quad (9)$$

$$W = \begin{cases} > 0 Group I \\ < 0 Group II \\ 0 Cannot classify \end{cases}$$

The last represents the value of the discriminant function at the border separating the two sets. If W >0, then the item is related to the first population, but if W<0, then the item is related to the second set. (W) is called the statistics of classification that was developed by Anderson in 1940.

2.4 The Methods used in Testing the Variables of the Discriminant Function

First: The Confidence Intervals of Roy-Bose

When the significance of the discriminant function is proven using Hotelling T2 and the variance table, a question is raised: which one of the significant variables led to the rejection of the hypothesis and accepting the alternative hypothesis? The confidence intervals of Roy-Bose are used to determine these significant variables in the classification, i.e. the variables that will be entered into the function. This method can be summarized with the following steps (Morrison, 1976):

1- Appointing the value of T (tab. T), where:

$$tab.T = \left[\frac{(n_1+n_2-2)}{n_1+n_1-P-1} tab.F\right]^{\frac{1}{2}} \quad (10)$$

2- The value of F table value is determined by referring to the relevant tables:

$tab.F = (\frac{\alpha}{2}, m, n_1+n_2-1)$

3- Finding the selection vector, which is symbolized by (a), which is non-zero vector, and the number of these vectors equals the number of the variables (P), as:

$a_1 = [1 \quad 0 \quad 0 \quad \dots]$ $\quad a_2 = [0 \quad 1 \quad 0 \quad \dots]$ ... $a_m = [0 \quad 0 \quad \dots \quad 1]$

4- After that the confidence intervals of Roy-Bose are found as follows:

$$\underline{a}_p'(\bar{\underline{X}}_1 - \bar{\underline{X}}_2) - \sqrt{\underline{a}_p' S_{\underline{a}_p}\left(\frac{n_1+n_2}{n_1 n_2}\right)} < \underline{a}_p'\gamma < \underline{a}_p'(\bar{\underline{X}}_1 - \bar{\underline{X}}_2) + \sqrt{\underline{a}_p' S_{\underline{a}_p}\left(\frac{n_1+n_2}{n_1 n_2}\right)} \quad (11)$$

The differences vector $(\bar{\underline{X}}_1 - \bar{\underline{X}}_2)$ is distributed normally with multivariable for both $\bar{X}_1, \bar{X}_2$, while the variance matrix and the pooled variance are calculated according to the equation (7):

5- Then the following matrix is written:

$C.I_{(L)} < \underline{a}_p'\gamma < C.I_{(u)}$

The decision is if the matrix contains zero, this means that the variable is not significant, i.e. the variable mathematical mean is not different in both of the sets. But, if the matrix didn't include zero, this means than the variable is significant.

Second: T test

After the significance of the discriminant function is selected, the significance of each variable in the discriminant should be tested because the number of the explanatory variables in the analysis could be enormous and the insignificant variables are excluded from the analysis and this leads to good results in the process of discrimination.

The t-test can be used to compare the mathematical means as it can be used to test these variables. The hypothesis of (t) test is:

$H_0: \Sigma_1 = \Sigma_2$

$H_1: \Sigma_1 \neq \Sigma_2$

On the other hand, the statistics of the test is shown in equation (6) and with the comparison of the value of the value of statistics with the t table value with a degree of freedom of (n_1+n_2-2) and a significance level of (α) will determine whether the

then the null hypothesis is rejected and the alternative hypothesis is |Cal.t > tab.t|variable is significant or insignificant. If accepted at a significance level of (α).

2.5 Fuzzy Discriminant Analysis (FDA)

The Fuzzy Discriminant Analysis is one of the recent statistical methods that is used in the field of data analysis. This method combines the principles of the traditional discriminant analysis and the fuzzy logic. This method aims at improving the accuracy of classification and achieving more flexibility in terms of dealing with uncertain or fuzzy data that is abundant in many domains like medicine, economy, marketing, and engineering (Qu, and Pei,2024).

The fuzzy discriminant analysis is a copy of the linear discriminant analysis (LDA) or the (qualitative discriminant analysis (QDA) that is applied in a fuzzy or a vague environment, where data is not specified or inaccurate instead of dealing with separate categories. In the fuzzy discriminant analysis, the samples belong to more than one category, with varying degrees, or "organic degrees" and this makes the analysis more flexible in dealing with the uncertainty of data (WU et al., 2023).

The traditional discriminant analysis depends on a set of features that are used in the data classification into various categories through discriminant functions that depend on the ratio between within-class variance and the variance between classes, but in the case of fuzzy data the borders between the data might be unclear and the analyst encounters a difficulty in classifying the points accurately. Here, comes the idea of fuzzy discriminant analysis, as the techniques of fuzzy logic are used to deal with the uncertainty. This method relies on transforming the data into fuzzy values and this allows the degree of each point to each category instead of putting them in one category only (Zhang, et al., 2024).

The fuzzy set theory tackles a type of uncertainty, which is vagueness that is related to normal languages. The theory was resented in 1965 by the Azerbaijani scientist Lutfi Zadah from California University and it is a concept that deals with data that represent vague and uncertain things like "very cold" In the same year, he published his research (the fuzzy sets), in which he stated the mathematical aspects of the fuzzy aspects theory as he paid attention to the complex systems and simplifying them using simple mathematical models(Qader, et al., 2023).

2.6 The fuzzy Set

It is a set with elements that possess a degree of membership, which is either full 100% or partial (less than 100% and more than 0%). The borders of this set is not acute and this concept contradicts with the traditional concept of the crisp set that has accurate borders (Klir et al., 1997).

2.7 The Crisp Set

It is a set of things that are featured with one characteristic that takes one of the two values: (1) when an element belongs to the set and (0) when the element doesn't belong to the set and it was called the crisp set to discriminate it from the fuzzy set in the concepts of the fuzzy sets. Suppose that we have the set (A), which is called as a function and it is named the distinguished function μ

$\mu_A: x \to \{0,1\}$

$$\mu_A(x) = \begin{cases} 0 & \text{if } x \notin A \\ 1 & \text{if } x \in A \end{cases}$$

For example, the word (warm) for the fuzzy set includes a vast domain for measuring the temperature of a certain area. When it says "the weather is warm" in a specific time, you can read the measurement degree "warm" in a certain place and it is different

from another place and so that depends on the location, season, day or night, etc… through this the fuzzy set can be defined by determining the temperature degree between (1 and 0), i.e. by placing a membership degree for the set (Klir et al., 1997).

If we have X as representing the inclusive set, then the fuzzy set A from X is a set of

$A = \{x, \mu_A(x) \qquad \forall x \epsilon X\}$

where; x is an element and $\mu_A(x)$ is a membership function of the element x to A, either the function of membership in the fuzzy set that is equivalent to the distinguished function in the crisp set except for that the fuzzy function can take any value between zero and one (AlDabbagh, 2003).

$0 \leq \mu_A(x) \leq 1$

2.8 Membership Degree

It is the amount of membership of a certain element to the fuzzy set and this degree is the degree that is restricted between zero and one (Kandel, 1986).

2.9 Membership Function

It is the function, through which the membership degree of a certain element to a certain fuzzy set is calculated. Each fuzzy set (A) identifies an inclusive set (X) as a function that corresponds to the characteristic function. This function is called a membership function and the function is symbolized by $\mu_A(x)$ and each x in the inclusive set (X) is given a value in the closed interval [0,1], as it distinguishes the membership degree of the element x in A. there are several types of the membership functions, they are (Klir et al., 1997):

1- The Triangular Membership Data

This function is characterized with three parameters (a, b and c), as shown in the following equation:

$$\mu_{A(x)} = \begin{cases} b\left(1 - \dfrac{|x-a|}{s}\right) & when \quad a - s \leq X \leq a + s \\ 0 & otherwise \end{cases}$$

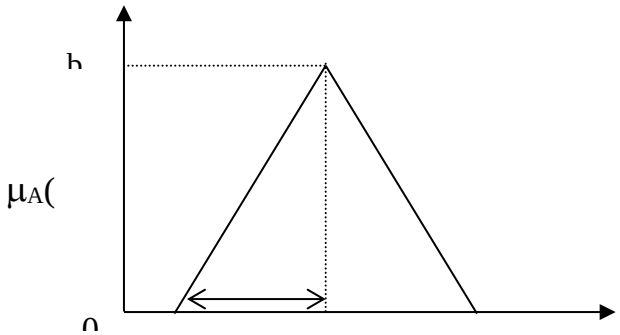

Figure (1): The triangular function

2- Trapezoidal Membership Function The formula of the :

$$\mu_{A(x)} = \begin{cases} \frac{a-x}{a-b} & : \quad a \le x \le b \\ 1 & : \quad b \le x \le c \\ \frac{d-x}{d-c} & : \quad c \le x \le d \\ 0 & : \quad otherwise \end{cases}$$

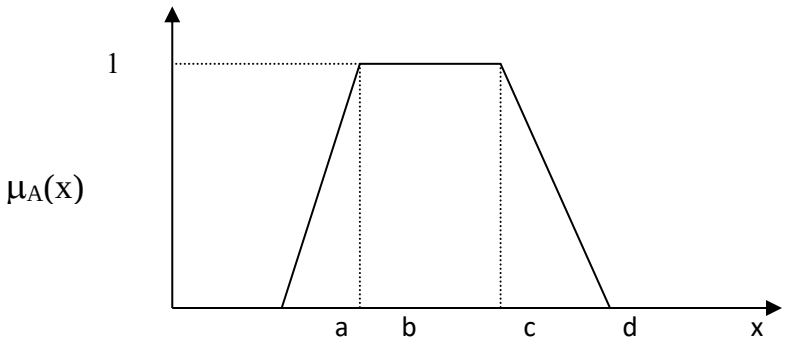

Figure (2): The Trapezoidal function

3- Bell-Shaped Membership Function also Called Gaussian Function

This function is used to identify the degree of membership in a fuzzy set. It adds to the curve of the membership a curved and a smooth shape as each element represents the degree of its membership to a certain set on a scale from 0 to 1 and this allows a more accurate and more flexible representation, as shown in figure (3).

$$\mu_A(X) = ce^{-\frac{(x-a)^2}{b}} \quad : \quad -\infty < x < \infty$$

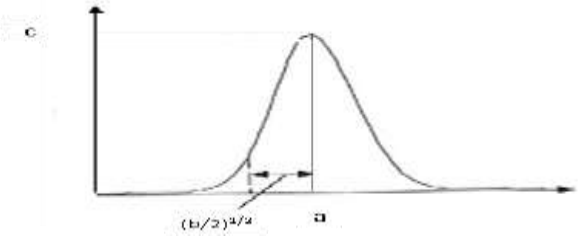

Figure (3): The Bell-shaped function

3. The Empirical Aspect

3.1 Description of the Research Sample and the Data Collection

In this aspect, two methods will be applied: The linear discriminant analysis and the fuzzy discriminant analysis to the data of the acute kidney failure patients. The results of the analysis and comparison between the two methods will be conducted depending on eleven explanatory variables that affect the disease (acute kidney failure). The data was obtained from the Ibn Sina Hospital in Mosul city in 2024 depending on the blood test (serum) of the kidney failure disease. Two samples were taken, the first represented the first set (the infected persons) and the second one stood for the (uninfected persons). The analysis was applied to the data using (R) package and (Matlab) package.

3.2 Variables Description

In this research eleven explanatory variables were depended upon as shown in table (2)

Table (2): Description of the explanatory variables

Variables	Variables description
X_1	Sex: (1) male, (2) female
X_2	Age
X_3	(1) smoker, (2) non-smoker
X_4	Urea in the blood
X_5	Creatinin in the blood
X_6	Calcium in the blood
X_7	Phosphorus in the blood
X_8	Alki phosphate
X_9	Glucose in the blood
X_{10}	Albumin in the blood
X_{11}	Total Peliropin

As for the response variable (y), it is a dummy variable as (0) represents the first set (infected) and (1) represents the second set (uninfected).

3.3 Data Statistical Analysis

After conducting the primary analysis in (Matlab) package, it was found that the number of the uninfected persons with acute kidney failure is (85) patients and with a percentage of (56.67%), while the infected persons were (65) unindividuals, with a percentage of (43.33%) as shown in table (3).

Table (3): shows the statistical data for the variables

Variable	Mean (0)	Mean (1)	StDev (0)	StDev (1)
x_1	1.38460	1.50590	0.50293	0.49029
x_2	0.65108	0.71953	15.15097	18.88220
x_3	1.56920	1.61180	0.49024	0.49904
x_4	0.69277	0.85129	12.01546	64.19588
x_5	0.68508	0.87976	0.39695	2.65029
x_6	0.60123	0.78929	1.21993	1.58184
x_7	0.60646	0.72553	1.10037	1.34026
x_8	0.61631	0.75506	84.06050	134.72561
x_9	0.70462	0.78929	48.97729	54.54110
x_{10}	0.70462	0.77753	0.66605	0.79066
x_{11}	0.91338	0.95424	0.50293	0.49029

3.4Test of the Means of To Sets

To test the existence of a difference between two means of sets, we use Wilks'sLambda Distribution test as shown in table (4):

Table (4): Wilks'sLambda Distribution test

Test of Function(s)	Wilks' Lambda	Chi-square	Df	Sig.
1	.472	107.237	10	.000

From table (4), it is noticed that sig. value is zero, which is lower than the significance level (0.01) and therefore, the hypothesis is rejected and the alternative hypothesis is accepted. This refers that there is a difference between the means of the two sets. Also, it is noticed that Wilks'Lambda statistics is close to the zero and eventually, the discrimination function has the ability to discriminate and classify the observations in accordance with the real population.

3.5 Finding the Estimate Values of the Function Parameters

The linear discriminant function can be determined through extracting the values of the function parameters depending on the following formula:

$$\alpha = S^{-1}(\bar{X}_1 - \bar{X}_2)$$

$$Z = 0.19132X_{i1} - 0.02534X_{i2} - 0.51980X_{i3} - 0.01555X_{i4} - 0.42405X_{i5} - 0.26021X_{i6} - 0.30921X_{i7} - 0.0097X_{i8} - 0.00545X_{i9} + 1.40889X_{i10} - 0.7242X_{i11}$$

3.6 Testing the Significance of the Linear Discriminant Function

In order to test the significance of the discriminant function and its ability to discriminate the populations, the two following tests can be applied:

First: Testing by using Hotelling T^2

In this test, the significance of difference will be identified depending on Hotelling T^2, as shown in equation (4) and D^2 will also be found through equation (5), as shown below:

$$D^2 = 5.0150$$

$$T^2 184.7191$$

Due to the difficulty of finding the table value and due to the unavailability of such tables, the value of table F can be taken directly from equation (6), as follows:

$$Ftest = \frac{85 + 65 - 11 - 1}{(85 + 65 - 2)11}(184.7191) = 15.658$$

Through the Hotelling T^2 and changing it into F test and comparing the calculated value of F with the table value of F at a significance level of (0.01), as:

tab.F=F(0.01,11,138)=2.4045

As the calculated F value is higher than the F table value, then we reject the null hypothesis and accept the alternative hypothesis. This refers that the linear discriminant function is efficient in discrimination, i.e. the difference are significant between the two sets, at least for one variable.

Second: Testing the Significance of the Discriminant Function using the Variance Analysis Table

It is also possible to test the significance of the discriminant function using the variance analysis table as shown below:

Table (5): variance analysis table

S.O.V.	d.f	SS	MS	Cal.F
Between X_s	11	6.2573	0.5688	15.652**
Within X_s	138	5.0150	0.03634	
Total	149	11.2723		

From table (5), it was noticed that cal.F=15.6528 is higher than Table F value (tab.F=2.4045) and this means that the discriminant function can be relied upon to classify and item to one of the set and also means that there is at least one variable that can be relied upon.

3.7 The Cut Point

For the purpose of classification, a cut point should be found that separates the two sets, depending on equation (11)

Cut Point= -4.4567

To obtain the classification equation (W) the cut point can be merged with the discriminant function, as follows:

$$W = 4.4567 - 0.02534X_{i2} - 0.01555X_{i4} - 0.42405X_{i5} + 0.26021X_{i6} - 0.30921X_{i7} - 0.0097X_{i8} + 0.00545X_{i9} + 1.40889X_{i10}$$

3.8 Methods used for Selecting the Variable of the Discriminant Function

Through testing the Hotelling T^2 and the variance analysis table, it is clear that the variables used for discrimination are significant variable, at least one variable, and can be used for the discrimination between two populations. But, these two tests didn't show which variable(s) are of important effect in discrimination and therefore, Roy-Bose method should be used in addition to t test to identify the variables with important impact in the discriminant function, which is a common method in the discriminant analysis.

1- The Confidence Intervals of Roy-Bose

To identify the significant variables, we should determine the confidence intervals to identify the significance or the insignificance of the explanatory variables, we apply equation (11). It was clear that all the explanatory variables are significant except for the variable (x_3) as the inequality included zero, as shown in the discriminant equation.

$$Z = 0.31150\,X_{i1} - 0.02560\,X_{i2} - 0.01570\,X_{i4} - 0.39499\,X_{i5} - 0.25777\,X_{i6} - 0.31135\,X_{i7} - 0.00979\,X_{i8} - 0.00578\,X_{i9} + 1.39065\,X_{i10} - 0.66893X_{i11}$$

For the purpose of classification, we find the cut point, that equals

Cut Point = -4.14002

$$W=4.14002 + 0.31150X_{i1} - 0.02560X_{i2} - 0.01570X_{i4} - 0.39499X_{i5} - 0.25777X_{i6} - 0.31135X_{i7} - 0.00979X_{i8} - 0.00578X_{i9} + 1.39065X_{i10} - 0.66893X_{i11}$$

2- t -Test

After determining the parameters of the discriminant function, the significance of the variables involved in the analysis and the ability of each variable to contribute in the discrimination process in order to omit or to exclude the unimportant variables in the classification. So, t- test was used to test the significance if the variables depending on equation (2):

Table (6): variance analysis table

x_i	Cal.t
x_1	2.9838
x_2	0.1845-
x_3	1.0629
x_4	0.2034-
x_5	4.4613-
x_6	3.9850
x_7	-6.2672
x_8	0.0641-
x_9	0.0566-
x_{10}	10.2545
x_{11}	0.7886-

From table (6), it is clear that the variables ($x_1, x_5, x_6, x_7, x_{10}$) are significant.

3.9 The Discriminant Function is Calculated Depending on the Variables that were Selected in the t-Test:

$$Z = 0.41026X_{i1} - 0.56422X_{i5} - 0.09957X_{i6} - 0.57392X_{i7} + 1.21231X_{i10}$$

For the purpose of classification, we find the cut point, which is:

Cut Point= 1.7061

$$W = -1.7061 + 0.41026X_{i1} - 0.56422X_{i5} - 0.09957X_{i6} - 0.57392X_{i7} + 1.21231X_{i10}$$

To classify any item depending on the function (W), we substitute the dependent variables that are related to this item in the equation and if W>0 that means it belongs to the first set (kidney failure patient), but if W<0 this means that the item belongs to the second set (uninfected person).

3.10 The Fuzzy Discriminant Function

Data is transformed into fuzzy values using the Gaussian Function using the Gaussian Function and this depends on identifying the fuzzy sets and determining the membership degree for each point and then the fuzzy discriminant functions are applied. Table (7) shows statistical data of the two sets after the data was fuzzed.

Table (7): The statistical data of the variables

Variable	Mean (0)	Mean (1)	StDev (0)	StDev (1)
x_1	0.2426	0.2580	0.0640	0.0624
x_2	0.7198	0.6514	0.2297	0.2880
x_3	0.6601	0.6429	0.0663	0.1051
x_4	0.8514	0.6928	0.1075	0.3077
x_5	0.799	0.6855	0.0527	0.3462
x_6	0.7888	0.6015	0.2334	0.2811
x_7	0.7244	0.6063	0.2292	0.2635
x_8	0.7550	0.6162	1.8952	0.2734
x_9	0.7896	0.7051	0.2292	0.2781
x_{10}	0.7778	0.7051	1.8952	0.2781
x_{11}	0.9535	0.9128	0.2281	0.1360

3.11 Testing the Means of the Two Sets

To test whether there are differences between the means of the two sets, Wilks's Lambda Distribution test was used, as shown in table (8):

Table (8) shows the results of Wilks's Lambda Distribution test

Test of Function(s)	Wilks' Lambda	Chi-square	Df	Sig.
1	.702	50.378	11	.000

From table (8) it is noticed that the value of sig. is zero, which is lower than the significance level (0.01) and so the hypothesis is rejected and this shows that there is a difference between the means of the two sets.

3.12 Finding the Estimated Values of the Function Parameters

The linear discriminant function can be calculated as the values of the function parameters can be determined depending on the following formula:

$$\alpha = S^{-1}(\bar{X}_1 - \bar{X}_2)$$

$$Z = \text{-}2.63729\, X_{i1} + 0.7589X_{i2} + 7041X_{i3} + 1.1442X_{i4} + 2.6424X_{i5} + 1.6019X_{i6} + 1.0837X_{i7} - 0.1334X_{i8} + 1.0036X_{i9} + 0.7622X_{i10} + 3.6991X_{i11}$$

3.13 Testing the Significance of the Linear Discriminant Function

To test the significance of the linear discriminant function and the its ability to discriminate between the research populations, the following two tests can be applied:

First: Using Hotelling T^2 Statistics

In this test the significance of the differences will be identified between the means using Hotelling T^2 as shown in equation (5). D^2 is found by equation (4), as follows:

$D^2 = 1.4800$

$T^2 = 54.5133$

Due to the difficulty of finding the table value and due to the unavailability of such tables, the value of table F can be taken directly from equation (6), as follows:

$Ftest = 3.1820$

Through the test Hotelling T^2 and changing it into F test and comparing the calculated values with the table F value at a significance level of (0.01):

tab.F=F(0.01,11,138)=2.375

As the calculated F value is higher than the F table value, then we reject the null hypothesis and accept the alternative hypothesis. This refers that the linear discriminant function is efficient in discrimination, i.e. the difference are significant between the two sets, at least for one variable.

Second: Testing the Significance of Discrimination Function using the Table of Variance Analysis

It is possible also to test the significance of the discriminant function through the a table for variance analysis, like in the case of table (9).

Table (9): Discriminant variance analysis

S.O.V.	d.f	SS	MS	Cal.F
Between X_s	11	0.545	0.0495	4.95^{**}
Within X_s	138	1.4800	0.010	
Total	149	2.025		

From table (9), it is observed that the value of cal.F=4.95 is higher than the F table value (tab.F=2.4045) and this means that the discriminant function can be relied upon to classify any item to one of the sets. It also means that there are, at least, one variable that can be relied upon.

For the purpose of classification, a separating point should be found that separates the two sets, depending on equation (11) as follows:

Cut Point= 8.7632

To obtain the classification equation (W), the cut point (CP) could be merged with the discriminant function (Z) and as follows:

$$W = -8.7632 + 1.1442X_{i4} + 2.6424X_{i5} + 1.6019X_{i6} + 1.0837X_{i7} - 0.1334X_{i8}$$

To classify any item depending on the function (W), we substitute the values of the dependent variables of this item in the equation. If W>0, this means that it belongs to the first set (infected with kidney failure), but if W<0, that will mean that the item belongs to the second set (uninfected with kidney failure), i.e.:

3.14 Methods used in Selecting the Variables for the Discriminant Function

Through the Hotelling T^2 test and the table of variance analysis, it was evident that the variables used for discrimination are significant variables and these variables can be used for the discrimination between two populations. But these two tests didn't demonstrate which of the significant variables are important in the discrimination and therefore, we will use the Roy-Bose method and t test to identify the variables that are effective in the discriminant function and this is a common method in the discriminant analysis.

First: The Confidence intervals of Roy-Brose

In order to identify the significant variables, we should find the confidence intervals of Roy-Brose and to identify whether the variables are significant or not we use equation (13). It was clear that all the variables are statistically significant as the inequality doesn't include zero except for the variable x_3, and this means that the variable has different mean in both sets.

The Discriminant Analysis using the Variables of Confidence Intervals of Roy-Brose

After identifying the important variables, we conduct the discriminant analysis for the variables of Roy-Brose.

$\alpha = S^{-1}(\bar{X}_1 - \bar{X}_2)$

2,75221

$Z = -\ 2{,}75221\ \ X_{i1} + 0.82781X_{i2} + 1.3398X_{i4} + 2.5027X_{i5} + 1.40551X_{i6} - 0.94550X_{i7} + 1.86265X_{i8} + 3.7706X_{i9} - 2.9584X_{i10} + 3.76381X_{i11}$

For the purpose of classification, we find the cut point that equals:

Cut Point= 274.0232

$W= \quad 274.0232 - 2.75221X_{i1} + 0.82781X_{i2} + 1.3398X_{i4} + 2.5027X_{i5} + 1.40551X_{i6} - 0.94550X_{i7} + 1.86265X_{i8} + 3.7706X_{i9} - 2.9584X_{i10} + 3.76381X_{i11}$

Second: t Test

After calculating the variables of the linear discriminant function, it is necessary to identify the significance of the variables that are involved in the analysis and the ability of each variable to contribute in the process of discrimination in order to omit or exclude the variables unnecessary in the classification. So, the t test was used to test the significance of the variables depending on equation (6). We determine the value of calculated t ($|Cal.t|$) and then compare it with the table t value at a significance level of (0.01) and a degrees of freedom of (n_1+n_2-2), as follows:

tab.t=$t_{0.005,\ 85+65-2}$=2.629

table (10) shows the results of t test for the difference between two mathematical means.

Table (10): The result of t test

Explanatory variables	$Cal.t$
x_1	1.46715-
x_2	1.62487
x_3	1.23436
x_4	4.42173

x_5	5.11617
x_6	4.47115
x_7	2.95800
x_8	3.67970
x_9	2.05069
x_{10}	1.73616
x_{11}	2.01370

From table (10) it was evident that the variables x_4, x_5, x_6, x_7, x_8 are significant

3.15 Calculating the Discriminant Function Depending on the Variables that were Selected by the t-Test

$Z = 2.04556X_{i4} + 1.13220X_{i5} + 2.57220X_{i6} + 1.74360X_{i7} + 0.47555X_{i8}$

For the purpose of classification, we find that the cut point equals:

Cut Point= 7.3849

$W = -7.3849 + 2.04556X_{i4} + 1.13220X_{i5} + 2.57220X_{i6} + 1.74360X_{i7} + 0.47555X_{i8}$

4.Conclusion

After making sure of the availability of the two methods of the traditional discriminant analysis and the fuzzy discriminant analysis, which is the nature of the data and the condition of the inequality of the means of the two sets and the significance of most of the affecting factors, the suitability of the discriminant function was reached using the methods of the traditional discriminant analysis and the fuzzy discriminant analysis for the data of the kidney patients, i.e., it is possible to use them in discriminating and classifying the new items (infected – not infected) according to the set of the explanatory variables.

When applying the confidence intervals by Roy-Bose to select the most important explanatory variables, the same results were obtained in the test of the variables that were entered to the traditional discriminant function and the fuzzy discriminant function as all the explanatory variables except the variable (x_3) which stood for (smoker, non-smoker). When applying the t test to the important variables in the traditional discriminant analysis method, the variables ($x_1, x_5, x_6, x_7, x_{10}$) were selected, but in the case of the fuzzy discriminant analysis the variables (x_4, x_5, x_6, x_7, x_8) were the ones that were chosen.

It was proven that the fuzzy discriminant method is betted that the traditional analysis and that was done through relying on the mean square error as its value was lower in the fuzzy discriminant analysis. And It is recommended not to use the fuzzy discriminant analysis method as it doesn't depend on the size and nature of the data, but only on the characteristics and type of that data and it also gave the leas error to the function.

References

1- Adebayo. O. P., Ogunjimi .O. & Ahmed. I.,(2024), " Application of MANOVA and Hotelling's T square on Academic Performance of University Students Based on Mode of Entry" , Iraqi Journal of Statistical Sciences, Vol. 21, No. 2, 2024, pp (1-8), https://stats.uomosul.edu.iq/article_185231.html .

2- Afifi , A.A. and Clark V. , (1984) , " Computer Aided Multivariate Analysis " Life time Leering Publications , Belmont , Catlifornia , U.S.A.

3- Ahmed, J., Zena Y., (2021), " Using fuzzy dynamic programming in finding the best solution for sales for Badoush cement factory stores", Iraqi Journal of Statistical Sciences, Vol. 18, No. 1, pp (66-73) , https://stats.mosuljournals.com/article_168380.html.

4- Arnold, S.F. (1981). " The Theory of Linear Models and Multivariate Analysis " , John Wiley and Sons , New York . https://doi.org/10.2307/2530188

5- Al- Rawi , Khashi Mahmoud , (1987) , " Introduction to Regression Analysis " , Printed by Dar Al-Kutub Foundation for Printing and Publishing , University of Mosul, 10.1088/1755-1315/1371/5/052035.

6- David G. Kupper, L (1978) , " Applied Regression Analysis and other multivariate methods" , The University of North Carolind and Chapel Hill, https://www.google.

7- Elkhouli, M,(2024), " The Implications of Discriminant Analysis Function in Classifying the Obesity of Childhood < 15 in Egypt", Iraqi Journal of Statistical Sciences, Vol. 21, No. 1, Pp (12-31), https://doi.org/10.33899/iqjoss.2024.0183229,

8- Kandel , A.(1986). Fuzzy Mathematical Techniques with Applications, Addison – Wesley Publishing Company , England . https://www.semanticscholar.org .

9- Karzan, F. , Bulent, c.m & Rizgar,M., (2024), "Classification of Circular Mass of Breast Cancer Using Artificial Neural Network vs. Discriminant Analysis in Medical Image Processing" , Iraqi Journal of Statistical Sciences, Vol. 21, No. 1, pp (46-58) , https://stats.mosuljournals.com/article_183231.html.

10- Klir, G.J. (1997). Fuzzy arithmetic with requisite constraints . Fuzzy sets and systems , 91(2) , 165- 175 , https://www.google.com.

11- Morrison , D.F. (1976), " Multivariate Analysis Aniversity of Pennsg Lvanid " ; New York .

12- Oladapo, O.J., Alabi ,O.O,& Ayinde, K. ,(2024)," Performance of Some Yang and Chang estimators in Logistic Regression Model Iraqi Journal of Statistical Sciences", Vol. 21, No. 1, PP (1-11) , https://doi.org/10.33899/iqjoss.2024.183228 .

13- Qader,H., Mahmood,M., Mrakhan,M.& Ramadan,R.,(2023), " Techniques to Restrict an Interval of a Lower Bound in Fuzzy Scheduling Problems" , Iraqi Journal of Statistical Sciences, Vol. 20, No. 1, Pp. (1-8), https://www.google.com.

14- Qu, L., & Pei, Y. (2024) . A Comprehensive Review on Discriminant Analysis for Addressing Challenges of Class-Level Limitations, Small Sample Size , and Robustness . Processes, 12(7) , 1382, https://doi.org/10.3390/pr12071382.

15- Wu, X., Fang , Y., Wu , B., & Liu , M . (2023). Application of nearinfrared spectroscopy and fuzzy improved null linear discriminant analysis for rapid discrimination of milk brands . Foods, 12 (21) . 3929, https://doi.org/10.3390/foods12213929.

16- Zhang , J., Wu, X., He, C., Wu, B., Zhang , S., &Sun , J. (2024) . Near- Infrared Spectroscopy Combined with Fuzzy Improved Direct Linear Discriminant Analysis for Nondestructive Discrimination of Chrysanthemum Tea Varieties . Foods , 13(10) , 1439, https://doi.org/10.3390/foods13101439

التحليل التمييزي المضبب مع التطبيق

رفل طلال سعدي حسين[1]، الاء عبد الستار حمودات[2]

[2،1]قسم الإحصاء والمعلوماتية، كلية علوم الحاسوب والرياضيات، جامعة الموصل، الموصل، العراق

الخلاصة: في هذا البحث تم استعمال اسلوبين في تصنيف البيانات و هما اسلوب التحليل التمييزي الخطي Linear Discriminant Analysis(LDA) وأسلوب التحليل التمييزي المضبب Fuzzy Discriminant Analysis(LDF) ، ويعد التحليل التمييزي طريقة لتصنيف مفردة ما إلى إحدى المجاميع معتمدا في ذلك على متغيرات ذات أهمية إحصائية في عملية التصنيف اذ يتم تصنيف المفردة الى المجتمع الصحيح من خلال مجموعة من المتغيرات . وفي بعض الأحيان تكون البيانات غير خطية او لا تتوزع توزيعاً طبيعياً او غير مستقرة فاستخدام المنطق المضبب يكون ناجحا لأنه لا يشترط ان تكون البيانات تتوزع توزيعاً طبيعياً او مستقرة ، ومن خلال النظرية المضببة يتم إيجاد نماذج لحل المسائل والتي لايمكن تحليليها باستخدام الطرائق الرياضية الصرفة، ويهدف البحث الى المقارنة بين اسلوبي التحليل التمييزي التقليدي (LDA) والتحليل التمييزي المضبب (LDF) وذلك لغرض تمهيد البيانات والتمييز بين الفئات المختلفة بدقة عالية ولتوفير تصنيفات دقيقة للمشاهدات ، كذلك تشخيص المتغيرات ذات الأهمية التصنيفية في الدالة التمييزية باستخدام طريقة حدود الثقة لـ Roy–Bose واختبار t ، اذ تم استعمال بيانات حقيقية لمجموعتين من مرضى الكلى (مصاب ، غير مصاب) بناءا على مجموعة من العوامل المؤثرة ومدى مساهمة كل عامل في التمييز الاشخاص او المفردات الجديدة وتوزيعها على احدى المجموعتين ، مما يؤدي الى التشخيص المبكر وتفادي تدهور الحالة الصحية للمريض ، وتم الحصول على النتائج باستعمال البرنامج الاحصائي (SPSS) وبرنامج (R) الى ان اسلوب التحليل التمييزي الضبابي هو الافضل لكونه يعطى اقل خطأ تصنيف للبيانات بالاعتماد على معيار الاحصائي Mean Squar Error (MSE) .

الكلمات المفتاحية : التحليل التمييزي ، T^2 Hotelling ، المنطق المضبب ، نقطة القطع