IRAQI Academic Scientific Journals

University of Kirkuk Journal For Administrative and Economic Science (13) المجلد (1): P-P 85-102

UKJAES

University of Kirkuk Journal For Administrative and Economic Science

ISSN 2222-2995
University of Kirkuk Journal For Administrative and Economic Science

Matrood Daham Owaid .& Al-Quraishi Osama Abdulazeez. **The use of linear discriminant analysis to determine the variables affecting cancer diseases.** ***University of Kirkuk Journal For Administrative and Economic Science*** **(2023) 13 (1): 85-102.**

The use of linear discriminant analysis to determine the variables affecting cancer diseases

Daham Owaid Matrood [1], Osama Abdulazeez Kadhim Al-Quraishi[2]

[1] *Northern Technical University, Nineveh, Iraq*
[2] *Ministry of Education, Baghdad, Iraq*

daham.stat@ntu.edu.iq
ousamastat@gmail.com

Abstract. The method of discriminant analysis is one of the multivariate statistical methods that used to deal with descriptive data, and it depends on building a discrimination function , which is a linear combination of a set of independent variables, and this function works to reduce the similarity in classification errors. Discriminate analysis is used to classify one or more items into correct groups with the lowest possible classification error. The linear discriminate function was introduced by (Fisher) in (1936), who suggested that the classification should be based on a linear combination of the discriminating variables. By maximizing group differences on the one hand and minimizing differences within groups on the other hand, this method is suitable when the classification error is as low as possible. Work was done on the empirical work method, in which we begin to observe the problem, then hypotheses were made for it then tested, that is, by selecting a random sample and generalizing its results to the community. Determine the factors affecting the incidence of lung cancer and bone cancer and classifying the data using the linear discrimination function in order to reach the best classification for some types of cancerous tumors on the basis of the probability criterion of the least possible classification error. The most important factor affecting the incidence of lung and bone cancer is the gender of the patient, then the period of stay in the hospital, and finally the work of the patient.

Keywords: Linear discriminant analysis, Cancer diseases, homogeneity, Determine, Eigen value.

Introduction

In recent years, there has been a noticeable increase in the number of infected people of Cancer which is a group of diseases characterized by aggressive cells (unlimited cell growth and division), and the ability of these dividing cells to invade and destroy nearby tissues, or move to distant tissues in a process called metastasis. These abilities are the characteristics of a malignant tumor, unlike a benign tumor, which is characterized by a specific growth and inability to invade and has no ability to move or transmit. Sometimes a benign tumor can develop into a malignant cancer. It is the non-normal growth of cells in one of the tissues of the body. It affects different types of organs, and the symptoms usually differ according to the affected organ or tissue. Many cancers can be prevented by avoiding exposure to common risk factors, such as tobacco smoke. A large proportion of cancers can also be treated with surgery, radiotherapy or chemotherapy, especially if they are detected in the early stages. Cancer affects all human life stages, even fetuses, but the risk of its developing increases with age. As cancer affects humans, some of its types affect both animals and plants. The method of discriminant analysis is one of the multivariate statistical methods that used to treat descriptive data, and it depends on building a discriminantion function, which is a linear combination of a set of independent variables, and this function works to reduce the similarity in classification errors. Discriminate analysis is used to classify one or more items into their correct groups with the lowest possible classification error. The linear discriminate function was introduced by (Fisher) in (1936), who suggested that the classification should be based on a linear combination of the discriminating variables. By maximizing group differences on the one hand and minimizing differences within groups on the other hand, this method is suitable when the classification error is as low as possible.

Research Importance

The importance of discriminant analysis comes in terms of its use in various fields of applied life, such as biomedical studies, education, agriculture, geography, earth science, and finally prediction . the research also provides a statistical model or a discriminatory function that has the ability to discrimination and separate between people into two groups . A patient for lung cancer and the other for bone cancer, then classify the new observations and distribute them to one of the two groups.

Research Methodology

Work was done on the empirical work method , in which we begin to observe the problem, then hypotheses were made for it and then tested , that is , by drawing a simple random sample and generalizing its results to the community.

Research Hypothesis

The assumed state of the spread of cancer is non-normal, and accordingly the hypothesis is:

H_0: The data follows a normal distribution.
H_1 : The data dose not follows a normal distribution.

Aim of the Research

The research aims to:

1. Shedding light on the method of discriminatory analysis as one of the methods of multivariate statistical analysis.
2. Determine the factors affecting the incidence of lung cancer and bone cancer.
3. Classifying the data using the linear discrimination function in order to reach the best classification for some types of cancerous tumors on the basis of the probability criterion of the least possible classification error.

Research community

people with lung cancer and bone cancer.

Research framework

spatially represents the research framework Baghdad Karkh, temporally represents the year 2022.

Data source:

Data were taken from a sample of 60 people from Yarmouk Hospital, and the sample is divided into 30 patients with lung cancer and 30 with bone cancer.

1.The Theoretical side:

1.1. Discriminant Analysis [4],[7]

Discriminant analysis is one of the important statistical methods in multivariate statistical analysis, as it is concerned with the issue of discrimination (differentiation) between two or more groups that are similar in many characteristics on the basis of several independent variables, and that the process of discrimination takes place after the formation of the discriminatory function (using the discriminatory function). Which is a linear combination of the independent variables, which is the one who predicts and classifies the sample items into two different groups, as the discriminatory function increases the difference between the average of the two groups, as the more there is a spacing between the average groups, the more efficient the discrimination and the least error classification with the least possible error In other words, the discriminatory function increases the degree of homogeneity between the vocabulary of one group and reduces the homogeneity between the two groups:

1.2. Types of Discriminant Analysis [1],[8]

a. Direct discriminate analysis:

Where the variables enter the analysis at once without giving any importance to any variable.

b. Hierarchical discriminant analysis:

The variables are entered according to the researcher's vision.

c. stepwise discriminate analysis:

The variables are entered for analysis according to a statistical standard that determines the priority of entering the variables into the model, where the variables are added to the discriminating functions one by one until we find that adding variables does not give a better discrimination.

1.3. Objectives of the discriminant analysis [10]

There are several objectives of discriminatory analysis, including:

- Create discriminant functions to separate or distinguish between the categories of the dependent variable.
- These functions maximize the difference between groups (categories of the dependent variable).
- The order of the variables that contribute significantly to discrimination or clarifying the differences between groups (categories of the dependent variable).
- Classification of new observations and their distribution to groups (categories of the dependent variable).

Reaching the lowest error rate of the characterization.

1.4. Conditions of Discriminative Analysis [12]

1. The independent variables are normally distributed for each population.
2. Equal The variance and covariance for the independent variables .
3. Choosing the sample at random.
4. There is a linear relationship between the variables used , and this can be investigated by drawing the shape of the spread for each pair of these variables.
5. There is no problem of multilinearity between the explanatory variables, that is, the variables are independent of each other.

1.5. Steps to Conduct a Discrimination [14],[11]

1. Finding the dependent variable: We determine the groups that we want to classify.

2. Choose the input variables the make up the discriminant equation:

The independent variables that make up the model are chosen by selecting the variables that have the highest value (F) and the lowest value (Wilks Lambda).

3. Standardization Discriminant coefficients: The standard discriminant coefficients are represented by the values of (a) shown in the following equation:

$$Y^* = a_1x_1 + a_2x_2 + \dots\dots\dots + a_px_p \qquad (1)$$

Since:
Y^*: the standard discriminant value.
x_p : standard discriminant variables.
a_p: the standard discriminant coefficients.
p: the number of standard discriminant variables that make up the discriminant equation.
The standard discriminatory equation is used to determine the importance of the variables in the formation of the discriminant function, as the variables whose absolute value is large, It contribute greatly to the formation of the discriminatory equation, and the sign of the standard discriminatory

coefficient means that the contribution of the ratio in discrimination is a positive or negative contribution.

1.6. Linear Discriminate Function [15],[2]

This method is one of the parametric methods and this type of function is used when the population under study has a multivariate normal distribution and the variances (the variance and covariance matrix ∑) are equal, with different averages μ1, μ2.

It is one of the simplest cases of discrimination, which requires the following conditions to be met:

1. The independent variables are assumed to be multivariately distributed normally.
2. The variances are equal for all totals (the variance and covariance matrices), that is, the null hypothesis is not rejected when the following hypothesis is tested:

$$H_0: \square_1 = \square_2 = \cdots = \square_K$$
$$H_1: \square_1 \neq \square_2 \neq \cdots \neq \square_K$$

Since:

Σ: variance and Covariance Matrices
K: the number of totals.

3. classifying observations the existing into groups (n1, n2 ...) Accurately.
4. There is no multi-linearity problem between the independent variables.

1.7. Linear Discriminate Function (LDF) –Tow Groups [16],[9]

The discrimination function is a model that can be formulated depending on the indicators of the sample whose vocabulary was chosen and placed in two different groups, and with this function we can test the item and determine its relevance to any group.

The steps for calculating the discrimination function between two groups are as follows:

1. Find the mean of each independent variable in each group.

a- first group:

$$\underline{x}_1{}^{(1)} = \sum_{i=1}^{n} \frac{x_{1i}{}^{(1)}}{n_1{}^{(1)}} \qquad \ldots..(3)$$

$$\underline{x}_2{}^{(1)} = \sum_{i=1}^{n} \frac{x_{2i}{}^{(1)}}{n_2{}^{(1)}} \qquad \ldots..(4)$$

$$\underline{x}_n{}^{(1)} = \sum_{i=1}^{n} \frac{x_{ni}{}^{(1)}}{n_n{}^{(1)}} \qquad \ldots..(5)$$

$$\underline{x}_i{}^{(1)} = \left(\frac{\underline{x}_1{}^{(1)}}{\underline{x}_2{}^{(1)}} \ldots \underline{x}_n{}^{(1)} \right) \qquad \ldots(6)$$

b- second group:

$$\underline{x_1}^{(2)} = \sum_{i=1}^{n} \frac{x_{1i}^{(2)}}{n_1^{(2)}} \quad \ldots..(7)$$

$$\underline{x_2}^{(2)} = \sum_{i=1}^{n} \frac{x_{2i}^{(2)}}{n_2^{(2)}} \quad \ldots..(8)$$

$$\underline{x_n}^{(2)} = \sum_{i=1}^{n} \frac{x_{ni}^{(2)}}{n_n^{(2)}} \quad \ldots..(9)$$

$$\underline{x_i}^{(2)} = \left(\frac{\underline{x_1}^{(1)}}{\underline{x_2}^{(1)}} \ldots \underline{x_n}^{(1)} \right) \quad \ldots(\ 10\ )$$

2. Finding the difference (distance) between the mean of each of the two variables in the group:

$$d_1 = \underline{x_1}^{(1)} - \underline{x_1}^{(2)} \quad \ldots..(11)$$

$$d_2 = \underline{x_2}^{(1)} - \underline{x_2}^{(2)} \quad \ldots..(12)$$

$$d_n = \underline{x_n}^{(1)} - \underline{x_n}^{(2)} \quad \ldots..(13)$$

$$d_i = \underline{x_i}^{(1)} - \underline{x_i}^{(2)} = \left[\underline{x_1}^{(1)} - \underline{x_1}^{(2)} \ \underline{x_2}^{(1)} - \underline{x_2}^{(2)} \ \vdots \ \underline{x_n}^{(1)} - \underline{x_n}^{(2)} \right]$$
$$= [d_1 \ d_2 \ \vdots \ d_n] \quad \ldots(\ 14\ )$$

3. Find the variance for each independent variable in each group:

a- first group:

$$s_i^{\,2} = \sum_{i=1}^{n} x_i^{\,2} - \frac{(\sum_{i=1}^{n} x_i)^2}{n_1} \quad \ldots\ldots(15)$$

b- second group:

$$s_i^{\,2} = \sum_{i=1}^{n} x_i^{\,2} - \frac{(\sum_{i=1}^{n} x_i)^2}{n_2} \quad \ldots\ldots(16)$$

4. Finding the variance and covariance within each group:

a- first group:

$$s_{ij} = \sum_{i=1}^{n} x_i x_j - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_j}{n_1} \quad \ldots\ldots(17)$$

b- second group:

$$s_{ij} = \sum_{i=1}^{n} x_i x_j - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_j}{n_2} \quad \ldots\ldots(18)$$

5. Finding the variance and the combined covariance (within the groups):

a- Built-in variance:

$$v_{ii} = \frac{s_i^{\,2(1)} + s_i^{\,2(2)}}{n_1 + n_2 - 2} \quad \ldots\ldots(19)$$

b- Built-in covariance:

$$v_{ij} = \frac{{s_{ij}}^{2(1)} + {s_{ij}}^{2(2)}}{n_1 + n_2 - 2} \quad \ldots\ldots(20)$$

And from equation No. (14), (15) we form the symmetric square matrix with the main diagonal representing the variances inside the totals and their other elements, the covariance as follows:

$$v = [v_{11}\, v_{12}\, \ldots\, v_{1k}\;\; v_{21}\, v_{22}\, \cdots\, v_{2k} \quad \vdots \quad \vdots \quad v_{k1}\, v_{k2}\, \cdots\, v_{kk}\;] \quad \ldots(21)$$

6. Finding the discriminate function as follows:

$$L = \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n \qquad \ldots.(22)$$

whereas:

α_i: discrimination coefficients

L: discriminate function

The α_i vector gives the highest percentage of discrimination between the two groups, and what is meant by the highest discrimination is that these coefficients make the differences in the values of the L function between the two groups much greater from the differences in the values of the function within the two groups, making it easy the process of discrimination between new vocabulary depending on the values of the independent variables by substitution in the values of those coefficients (discrimination function).

The percentage of differences within the two groups is denoted by the symbol λ, then:

$$\lambda = \frac{between\ group\ varation}{witH\ in\ group\ varation} \qquad (23)$$

After that, we choose the coefficients that make the difference ratio λ as large as possible, where we estimate the discrimination function by maximizing the ratio λ by deriving it partially and setting it equal to zero, so the normal equations to find the coefficients α_i is:

$$[v_{11}\, v_{12}\, \ldots\, v_{1k}\;\; v_{21}\, v_{22}\, \cdots\, v_{2k} \quad \vdots \quad \vdots \quad v_{k1}\, v_{k2}\, \cdots\, v_{kk}\;][\alpha_1\, \alpha_2 \vdots \alpha_k] = [d_1\, d_2 \vdots d_k] \quad (24)$$

$$\alpha = v^{-1}\, d \qquad (25)$$

1.8. The relative importance of the explanatory (independent) variables.[17],[13]

The discriminant analysis is to determine the relative importance of the independent variables affecting the discriminate process, through the following equation:

$$\alpha^*{}_i = \alpha_i \sqrt{v_{ii}} \qquad (26)$$

When comparing the absolute value of $\alpha^*{}_i$, the largest value means that the corresponding variable xi is the most important variable that has the ability to discriminant between the two groups and the second largest value of $\alpha^*{}_i$ means that the corresponding variable xi is the second most important variable that has the ability to Distinction and so on until the last variable has the ability to distinguish between the two groups.

1.9. Statistical Analysis of the discriminant Function [5],[12],[4],[3]

Before defining the discriminantion function, some tests must be done, including:

First: To test the ability of the discriminant function to discriminate .When we want to discrimination between two populations, we can test the hypothesis that the means of the two groups are equal:

$H_0: \mu_1 = \mu_2$

$$H_1: \mu_1 \neq \mu_2$$

Several measures can be used to test the above hypothesis:

$$Hotelling\ T^2\ Test:$$

It is used in the case of discrimination between two groups, as the test statistic is as follows:

$$T^2 = \frac{n_1 n_2}{n_1 + n_2} D^2 \qquad (27)$$

Since:

$$D^2 = \left(\underline{X}_1 - \underline{X}_2\right)' S^{-1} \left(\underline{X}_1 - \underline{X}_2\right)$$

Because there are no tables available for this purpose, due to the difficulty of extracting the tabular value, the tabular value (F) can be obtained directly, which was developed by (Rao) in 1952, where the (F) test is directly related to the Hoteling T^2 statistic, as its formula is as follows:

$$F = \frac{n_1 + n_2 - P - 1}{(n_1 + n_2 - 2)P} T^2 \qquad (28)$$

Degree of freedom $(n_1 + n_2 - P - 1, P)$

We reject H_0below the level of significance Fα if:

$$F_C > F_{\alpha(n_1+n_2-P-1,P)}$$

Wilks-scale:

To test the existence of a linear relationship between the variables, we test the following hypothesis:

H_0:there is no linear relationship between the variables.

H_1: there is linear relationship between the variables.

The test statistics are:

$$\square = \frac{|W|}{|T|} = \frac{|W|}{|W + B|} \qquad (29)$$

As:

W: represents the variance and covariance matrix within the groups.

B: represents the variance and covariance matrix between groups.

T: represents the variance matrix and the covariance of the groups.

The value of ($\square$) ranges between zero and one. If its value is equal to or close to one, this indicates that the averages of the groups are equal, that is, there is no discrimination between the groups, but if its value approaches zero, this indicates the strength of discrimination.

The above formula is distributed approximately χ^2 with a degree of freedom of p (k-1), so if the P-value is less than 0.05, we reject H_0 meaning that there is a linear relationship between the variables, meaning that the discriminatory function has the ability to distinguish.

Eigen value:

The eigenvalues (characteristic roots) are used to determine the extent of the ability of the discriminant function between groups, as the high value of the characteristic roots is an indication of the function's ability to discriminant

between groups.

second: Test equality matrix the variance and covariance

The equality of the variance and covariance matrix is one of the important conditions that must be met in order to be able to apply the linear discriminant function, and that the hypothesis is:

$$H_0: \square_1 = \square_2 = \ldots = \square_K$$
$$H_1: at\ least\ tow\ of\ them\ are\ not\ equal$$

The Barttlete test is one of the tests that are applied to verify from condition of homogeneity of variance, where the above hypothesis is tested. The formula of the test statistics is as follows:

$$M = Ln|S| \sum_{i=1}^{K} n_i - \sum_{i=1}^{K} Ln|S_i| \quad (30)$$

$$S = \frac{\sum_{i=1}^{K} n_i S_i}{\sum_{i=1}^{K} n_i} \quad (31)$$

Since:
S: Estimated variance and covariance matrix.

Si: the sample variance which is an unbiased estimate of Σ i since (i =1, 2, ….k)

n i: the degree of freedom of sample i .

k: the number of groups.

And Box proved that when M is multiplied by the constant C^{-1}, which equals:

$$C^{-1} = 1 - \frac{2P^2 + 3(P-1)}{6(P+1)(K-1)} \left[\sum_{i=1}^{K} \frac{1}{n_i} - \frac{1}{\sum n_i} \right] \quad (32)$$

Since:

P: The number of explanatory variables.

Then we reach to a scale distributed to χ^2 in degrees of freedom.

$\frac{1}{2}(K-1)(P-1)$ when ni is large.

So:

$$Box\ s'M = MC^{-1} \sim \chi^2_{\frac{1}{2}(K-1)(P-1)} \quad (33)$$

1.10. Classification [3],[10]

Before starting the classification process, we must refer to the separation point which is written as follows:

$$L = \frac{\underline{L}_1 + \underline{L}_2}{2} \quad \ldots.(34 \quad)$$

L: separation point

$\underline{L}_1$: the average of the discriminant values of the first group.

$\underline{L}_2$: the average of the discriminant values of the second group.

Therefore, the classification process is the next process after the formation of the discriminant function and the use of its ability to discriminant using the separation point.

Observations are classified into the first group if they are $\hat{L} > L$

Observations are classified into the second group if $\hat{L} < L$

The group is randomly classified into group one and group two if: $\hat{L} = L$

1.11: Classification error:[6],[9]

Classification error is defined as the probability of classifying an item (observation) to group (i) when in fact it belongs to group (j) and vice versa, and this incorrect classification occurs when determining the separating point between the two groups.

Types of classification error

First: the apparent classification error: It is calculated from the following classification table

Group	Follower of first group	Follower of second group	total
Group 1	n_{11}	n_{12}	n_1
Group2	n_{21}	n_{22}	n_2

n_{11}:The number of items from the first group that were classified in the same group and thus were correctly classified.

n_{12}: The number of items from the first group that were classified as an error in the first group.

n_{21}: The number of items that originally belonged to the second group and were classified as an error in the first group.

n_{22}: The number of items from the second group that were classified in the same group and thus were correctly classified.

The apparent error is calculated as follows:

$p_{12} = \frac{n_{12}}{n_1}$

p_{12}: Percentage of items belonging to the first group and classified as an error to the second.

$$p_{21} = \frac{n_{21}}{n_2}$$

p_{21}: Percentage of items belonging to the second group and classified as an error to the second.

The apparent error rate can be calculated using the equation:

$$\frac{n_{12} + n_{21}}{n_1 + n_2}$$

The real error: It represents the percentage of wrong classification in population:

$$p_{21} = p_{12} = F(\frac{-\sqrt{D^2}}{2})$$

In that F is the standard distribution function,

D is a MAHALONOBIS statistic

The value between the brackets is calculated and the corresponding probability is calculated from the standard normal distribution table. The closer the probability is to zero, the more it indicates a class, and the lower the description error, and thus the function's ability to discriminant and classify. But if the probability is close to one, it indicates a high description error and the ability of the function to distinguish and classify.

2. The practical side

2.1. Definition of variables

A- The dependent variable (Y), which represents the type of disease:

Bone cancer = 1 , Lung cancer = 2

B- The explanatory variables represent the following:

1- X_1= gender: (male = 1, female = 2)

2- X_2 = age groups (5 to less than 20), (20 to less than 35), (35 to less than 50), (50 to less than 65), (65 to less than 80)

3- X_3 = patient's profession:

(Child = 1, housewife or earner = 2, disabled = 3, retired = 4, employee = 5

4- X_4 = Patient's discharge status: (Death = 1, Referral = 2, Discharged = 3,Improvement =4)

5- X_5 = the patient's stay in the hospital:

(One month = 1, two months = 2, 3 months = 3, 4 months, 5 months and more).

2.2. Some important tests

Some tests must be carried out to verify the conditions for applying the discriminatory and agency analysis:

Normal distribution test for each population:

We test the data to know whether the independent variables of the two groups for cancer diseases are normally distributed or not using (Kolmogorov-Smirnov) and (Shapiro-Wilk W test for normal data) and according to the statistical program stata and according to the following hypothesis:

H_0: Normal distribution trace data.

H_1: Normal distribution trace no data .

Table 1: data test results for normal distribution

Variables	Kolmogorov-Smirnov		Shapiro-Wilk W test for normal data	
	Statistic	Sig.	statistic	Sig.
gender (X1)	0.37	.000	0.999	0.08
Eage(X2)	0.14	.150	0.995	0.28
Work(X3)	0.16	.087	0.988	0.67
M(X4)	0.15	.14	0.995	0.26
Time(X5)	0.15	0.104	0.986	0.74

In Table (1) the results of the Shapiro-Wilk W test for normal values showed that the level of significance for all variables is greater than (0.05), and accordingly we accept the null hypothesis that the data follow a normal distribution, the results of (Kolmogorov-Smirnov showed that the level of significance for all The variables are greater than 0.05)) and therefore we accept the null hypothesis that the data follow the normal distribution with the exception of the sex variable, and since the data size is 60 observations, we can consider that the data of the sex variable are close to the normal distribution according to the theory (central limit) [28].

Ensure that there are no outliers:

The box plot is used of the independent variables to detect the presence of outliers and conducted test the Mahalanobis under the hypothesis following:

H_0: no outliers.
H_1: there are outliers.

Table 2: shows the results of the Mahanlops values

MAH	MAH	MAH	MAH	MAH	MAH
6.42343	4.91273	4.03114	6.50326	4.93266	3.54037
2.32455	8.19667	5.58060	4.11057	7.77405	2.67108
1.48242	1.39622	4.49370	8.61891	7.02512	1.63730
4.29641	4.55446	7.26156	8.77499	3.54037	1.63730
2.83619	4.83863	6.73635	6.72396	6.53775	2.40870
6.66870	4.49370	4.77766	1.52905	3.54037	4.17973
10.06442	3.72577	6.43576	3.43930	6.86162	2.63594
4.91273	9.20769	4.61174	6.16376	6.09954	3.80312
5.21906	5.38625	1.55661	6.42294	3.44790	2.63594
3.80748	3.72371	7.73590	5.97053	4.81331	5.32830

By reviewing all the values, we find that each of these values is less than the tabular value of χ^2 at a degree of freedom (5-1) (2-1) and with a significance level of 0.05, which equals (18.4). Accordingly, we accept the null hypothesis that there are no outliers between All data related to all explanatory variables.

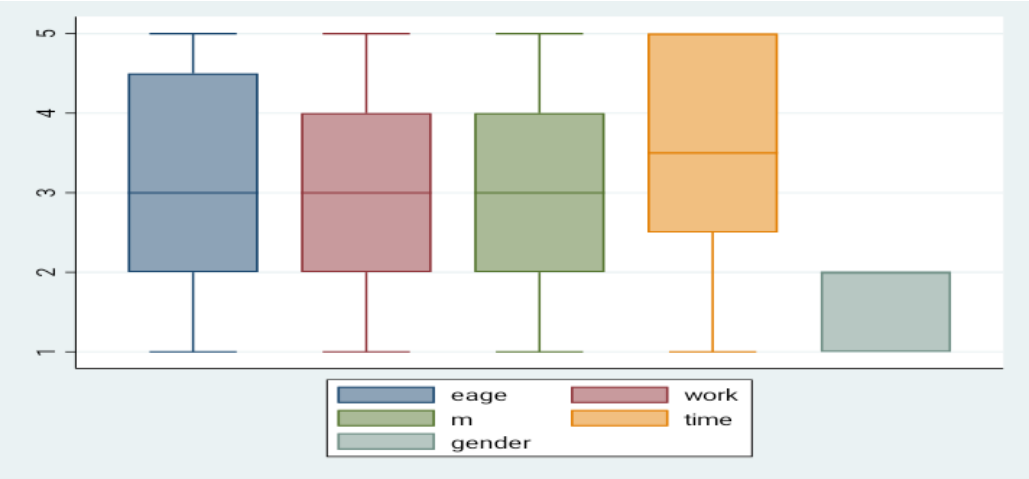

Figure 1: box diagram

By looking at the above figure, we do not see outliers outside the box, which confirms the Mahalanobis test.

Ensure that there is no polylinearity between the explanatory variables:

We make sure that there is no high correlation between the explanatory variables whose presence affects the degree of accuracy of the results. We calculate the tolerance coefficient as:

Tolerance =1 – R2xi, others

R2xi, others: others: represents the square of the multiple correlation coefficient between the dependent variable and the rest of the explanatory variables.

Then we extract the value of VIF (Variance Inflation Factor) for each of the explanatory variables as:

Table 3: Correlation coefficient test

Variables	Collinearity Statistics	
	Tolerance	VIF
gender (X1)	.866	1.15
eage (X2)	.914	1.09
Work(X3)	.843	1.19
m (X4)	.902	1.11
(time (X5)	.914	1.11

It is clear from **Table (3)** that the value of VIF for all variables is less than 0.05. It can be concluded that there is no problem of multilinearity among the independent variables.

Variance homogeneity test

To find out the homogeneity of the group members, we test the following hypothesis:

H_0: $\sigma_1 = \sigma_2$

H_1: $\sigma_1 \neq \sigma_2$

A test can be used (Box's M)

Table 4: shows the coefficient of determination

Log Determinants		
Y	**Rank**	**Log Determinant**
lung cancer	3	-1.174
bone cancer	3	-1.438
Pooled within-groups	3	-1.176

Table (5) shows the test for homogeneity of variances between groups

Box's M		7.537
F	Approx.	1.185
	df1	6
	df2	24373.132
	Sig.	.311

According to **Table (4)** the results indicated that the log determinant values are approximately equal for the totals, which indicates the homogeneity of the individual groups. From the test for equal variances in Table (5), it was found that the value of (Box's M = 7.537) with a significant level of (0.311) and this indicates the homogeneity of the variance between groups.

3.3. Steps to conduct a differentiation analysis:

Choosing the variables that make up the discriminant equation:

Was tested The significance of the variables to find out the importance of each variable individually and its impact on constructing the linear discriminant function, and the results were as in the table: (6).

Table 6: shows the significance of the discriminant function variables test.

Tests of Equality of Group Means					
	Wilks' Lambda	**F**	**df1**	**df2**	**Sig.**
(X1) gender	.747	19.5	1	58	.000
(X2) eage	.986	.820	1	58	.369
(X3) work	.877	15.473	1	58	.000
(X4) m	.734	121.05	1	267	.646
(X5) time	.996	.213	1	58	.000

Through **Table (6)** it is clear that the value of (P-Value = 0.000) is less than and greater than 0.05. We conclude that only three variables have an impact and importance in the formation and construction of the discriminant function, which is X1, X3, X5.

Table 7: shows the variables included in the analysis

Variables in the Analysis				
	Step	**Tolerance**	**F to Remove**	**Wilks' Lambda**
1	time(X5)	1.000	21.052	
2	time(X5)	.977	17.258	.747
	gender(X1)	.997	15.889	.734
3	time(X5)	.986	12.214	.632
	gender(X1)	.995	12.656	.636
	Work(x3)	.987	5.898	.574

We note from the above table that the variable x2 and x4 were excluded, and three variables were kept.

Standard discriminatory coefficients

Table 8: shows the coefficients of the standard discriminant function

Standardized Canonical Discriminate Function Coefficients	
variables	**function**
Time (x5)	0.615
Gender(x1)	0.621
Work(x3)	0.448

Table (8) shows the Canonical correlation between the discriminatory functions and each of the independent variables that were entered in the discriminatory analysis in standard units of measurement, i.e., determines the relative importance of the explanatory variables in estimating the independent variable. We note that the gender variable whose standard value is (.621) has the greatest effect in the incidence of lung cancer and bone cancer, followed by the time variable and finally the work variable.

Y= 0.621x1+0.615x5+0.448x3

Table 9: Relative importance

Canonical structure	
Variable	**function**
Time x5	.63
Gender x1	.60
Work	.53

Table (9) shows the relative importance, as the variable x5 contributes a greater percentage to the process of distinguishing between the two groups by percent. 63% is followed by the variable x1 with a percentage of 60. % And finally, a variable x3 with a percentage of 53 %.

3-2: Testing the ability of the discriminating function to discriminate

$$H_0: \mu_1 = \mu_2$$
$$H_1: \mu_1 \neq \mu_2$$

Wilks' Lambda:

Table 11: Test the significance of the discriminant function

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1	.32	37.046	3	.000

Note that the value of lambda, which is closer to zero and this indicates the high ability of the function to discrimination, and that the value of sig = .000 and thus reject H_0 and accept H_1 and that the function has the ability to discrimination.

Hotelling-lawely test:

Table 12: test the significance of the discriminant function

Test of Function(s)	Hotelling T2	Hotelling F	SIg
1	0.92	17.29	.0000

We see that the value of sig = .000 for the Hotelling test and thus reject $H_0$and accept $H_1$and that the function has the ability to discrimination.

Eigenvalues:

Table 13: shows the eigenvalues

Eigen values				
Function	**Eigen value**	**% Of Variance**	**Cumulative %**	**Canonical Correlation**
first	1.02	100	100	.693

Table (13) shows the value of the Eigen values for the discriminatory function, as it was (1.02), which indicates that the discriminatory function has a high ability to distinguish, as the value of Eigen values is greater than the correct one. As for the canonical correlation, it related (0.693) and this indicates that the quality of reconciling the discriminatory function.

2.4. Classification of the linear discriminant function

Table 13: shows the classification of the linear function

True y	**Classified**		**Total**
	2	**1**	
1	26	4	30
	86.67%	13.33%	100
2	6	24	30
	20%	80%	100

Table (13) shows the accuracy of the final results for classifying the linear function, as it turns out that (26) cases of the first group, with a rate of 86.7%, were classified correctly, and that (4) cases, with a rate of 13.3%, were classified incorrectly.

At the same time, it was found that 24 cases of the second group, with 80%, were correctly classified, and 6 cases, with 20%, were incorrectly classified.

As a general result, the results indicated that 83.4% of the cases in both groups were correctly classified, and this indicates a high quality of the classification results.

Conclusions

1. The data for the independent variables are normally distributed.
2. The averages of the group of people with lung cancer and those with bone cancer are not equal.
3. The variance matrix is equal to the covariance between the two groups.
4. Appropriateness method of the linear discriminant analysis and it can be used to discrimination groups and classify new vocabulary into people with lung cancer and with bone cancer according to a set of independent factors.
5. The variables involved in the formation of the discriminatory function are the patient's gender, occupation and period of stay in the hospital.
6. The discriminant function for separating and discrimination between the two groups with standard coefficients is:

   $Y = 0.621x_1 + 0.615x_5 + 0.448x_3$

7. The most important factor affecting the incidence of lung and bone cancer is the gender of the patient, then the period of stay in the hospital, and finally the work of the patient

8. The relative importance appeared that the variable $x_5$contributes a greater percentage to the process of discrimination between the two groups, at a rate of 0.63 % Is followed by the variable $x_1$with a percentage of 0.60 % and finally, a variable x_3 with a percentage of 53 %.
9. The results showed that 83.4% of the cases in both groups were correctly classified, and this indicates a high quality of the classification results.

Recommendations

1. We recommend the use of other discriminant functions (Kernel function, rank function, logistic function, quadratic function, Naif Bayes classification).
2. Expanding the number of variables in the study to include (smoking, heredity, alcohol consumption, intense exposure to harmful rays, viral or bacterial infection, ...) through a comprehensive form with all the information.
3. We emphasize when analyzing any problem, the integrity of the data and its compatibility with the basic conditions (normal distribution, anomalous values, homogeneity, independence,....) of the statistical method used to arrive at a probabilistic model with the least possible error.
4. It is necessary to develop statistics centers in hospitals, as well as new information for infected patients.

المستخلص: يعد أسلوب التحليل التمييزي من الاساليب الاحصائية متعددة المتغيرات التي تستخدم لمعالجة البيانات الوصفيه، ويعتمد على بناء دالة تمايز وهي عبارة عن توليفة خطية لمجموعة من المتغيرات المستقلة وهذه الدالة تعمل على تقليل التشابه في اخطا التصنيف ، يستعمل تحليل التمايز (Discriminate Analysis) لتصنيف مفردة واحدة أو أكثر الى مجموعاتها الصحيحة باقل خطا تصنيف ممكن قدمت دالة التمييز الخطية (Linear Discriminate) من قبل (Fisher) عام (1936)، الذي أقترح بأن التصنيف يجب أن يستند إلى تركيبة خطية للمتغيرات التمييزية من خلال تعظيم فروق المجموعات من جهة وتقليل التباينات داخل المجموعات من جهة أخرى وهذه الطريقة مناسبة عندما يكون خطأ التصنيف فيها اقل ما يمكن، تم العمل على منهج العمل الإستقرائي وفيه نبدأ بملاحظة المشكلة تم وضع الفروض لها ومن ثم اختبارها، اي من خلال سحب عينة عشوائية بسيطة حجمها 60 مريض بمرض السرطان (سرطان الرئة، سرطان العظام) وتعميم نتائجها على المجتمع. وهدف البحث إلى تحديد العوامل المؤثرة في الاصابة بأمراض سرطان الرئة وسرطان العظام وتصنيف البيانات باستعمال دالة التمييز الخطية بهدف الوصول إلى أفضل تصنيف لبعض أنواع الأورام السرطانية على أساس معيار احتمال أقل خطأ تصنيفي ممكن، وتوصل النتائج الى المتغيرات الداخلة في تكوين الدالة التمييزية هي جنس المريض والمهنة وفترة بقاءه في المستشفى حيث ظهر اكثر العوامل المؤثرة واهمها على الاصابة بمرض سرطان الرئة والعظام هو جنس المريض ثم فترة بقاءه بالمستشفى واخيرا عمل المريض.

References

1. Abbas, Lazim Muhammad,&Others, Classification according to some special physical abilities, complex skill performance and anthropometric measurements of basketball players, Published research, College of Physical Education and Sports Sciences, University of Al-Qadisiyah,2018.
2. Abdul Karim, Anwar diyaa, The use of statistical discriminant methods for the diagnosis of some heart diseases, Published research, Kirkuk University Journal for Scientific Studies, Volume (1), Issue (2), College of Science, University of Kirkuk,2017.
3. Abdul Razzaq, Aseel, Using the linear discriminant function to classify thalassemia patients in Kirkuk General Hospital, Published research, Journal of Management and Economics, Volume (11),Issue 78, Al-Mustansiriya University,2019.

4. Al.Katib, Muhammad Osama, The linear characteristic function in the case of more than two sets, Published research, Journal of Education and Science, Volume (23), Issue (4), Department of Computer Systems, Technical Institute, Nineveh,2018.
5. Al-Jaouni, Fredo Ghanem, Adnan, Multivariate statistical analysis (discriminant analysis) in the characterization and distribution of families within the socio-economic structure of society, Published research, Damascus University Journal of Economic and Legal Sciences, Volume (23), Issue (2),2018.
6. Fadl Almul, Suleiman, Ali Absher, Comparison between discriminant analysis, binary logistic model and neural network models in classifying observations, PhD thesis in statistics, College of Graduate Studies, Sudan University of Science and Technology,2019.
7. Gbara, Azhar Kazim, Analysis of multi response data for diagnosing eye diseases using the discriminant function and logistic regression (acomparative study), Master's thesis, College of Administration and Economics, Al-Mustansiriya University,2016.
8. Hammoudat, Alaa Abdel Sattar, The discriminant function and methods for determining its variables, Master's Thesis in Mathematics, University of Mosul,2019.
9. Hardle W., "Multivariate statistics", Printed on acid – free paper, p.227, (2007).
10. Huberty, C. J., "Applied discriminant analysis", New York, John Wiley & Sons, Inc, (1994).
11. Hunaiti, Dokhi & eta, Distinguishing the poor from the non-poor families in the remote areas of the Southern Jordan region, Journal of Economic Development and Policies, Volume (7), Issue (1),2017.
12. Izenman, A.J., "Modern Multivariate Statistical Techniques, Regression, Classification, and Manifold Learning", New York, Springer Science, Business Media, LLC, (2008).
13. Krieng kitbumrungrat, "Comparison logistic Regression and Discriminant Analysis in Classification groups for breast Cancer", Faculty of Information Technology, Rangsit University, Thailand, (2012).
14. Morrison, Donald F., "Multivariate statistical Method", second Edition, Tokyo, London. Mexico, New Delhi, (1976).
15. Nie, Norman H., et al., "Statistical Package for the Social Science", 2nd ed. (N.Y., McGraw-Hill Book Company), (1975).
16. Rencher, A. C., "Methods of Multivariate Analysis", John Wiley & sons, New York, USA, (1995).
17. Saleh, Samira Muhammad, Using the discriminant analysis (classification) to determine the most important factors affecting the dropout and failure of students at all academic levels, Master's Thesis in Statistics, College of Administration and Economics, University of Sulaymaniyah,2018.