The Journal of Administration & Economics / year 42/No 119/2019
ISSN : 1813-6729 http://doi.org/10.31272/JAE.42.2019.119.4

# "The Use of Mahalnobis Statistic in The Linear Discriminant Analysis between two Groups"

Dr. Sabah F. Abdulhussein*

**المستخلص:**

يهدف هذا البحث إلى تسليط الضوء على أهمية إحصاء مهالتوبيس($D^2$) في التمييز ما بين مجموعتين . ولتحقيق ذلك صاغ الباحث قاعدة للتمييز مشتقة من إحصاءةمهالتوبيس($D^2$) أسماها ($R_{D^2}$) وتتوزع توزيع ($F$).

كما أنه استخدم ($D^2$) لبناء دالة تمييز خطية ($L.D.F$) ولتقييم تصنيفها لمفردات عينة عشوائية.
إستفاد الباحث من التشابه وثبات نسبة التباين ($F$)في التحليل التمييزي الخطي ($L.D.A$)وتحليل الانحدار الخطي ($L.R.A$) وكذلك من العلاقة ما بين إحصاءةمهالتوبيس($D^2$)وكل من إحصاءة فشر ($F$) وإحصاءةهوتلن($T^2$).

وقد طبق الباحث إحصاءةمهالتوبيس($D^2$) ونسبة التباين المشتقة منها ($R_{D^2}$) على عينة عشوائية من المرضى المراجعين لمدينة الطب.

الكلمات المفتاحية: إحصاءة مهالنوبيس ($D^2$) ، ($L.D.F$) ، ($L.D.A$) ، ($L.R.A$)،إحصاءة فشر ($F$) ، إحصاءة هوتلن($T^2$).

## Abstract :

The purpose of this research is to highlight the importance of Mahalnobis has statistic ($D^2$) in the discrimination between two groups. To do that, the researcher formulated a rule of distinction derived from ($D^2$) and distributed ($F$) called ($R_{D^2}$). Also ($D^2$) has you used to construct a linear discriminant function ($L.D.F$) and evaluate its classification of random sample objects.

The researcher has of on advantage of similarity and the invariant Ratio of variance in the linear discriminant analysis ($L.D.A$) and linear regression analysis ($L.R.A$). As well as the relation between ($D^2$) and each of ($F$) and ($T^2$).

($D^2$) and ($R_{D^2}$) have been applied on random sample taken from patients in medical city.

Key-words: Mahalnobis Statistic ($D^2$), $L.D.F$, $L.D.A$, $L.R.A$ , Fisher Statistic ($F$) , Hotteling Statistic ($T^2$).

* معهد الادارة / الرصافة .
**مقبول للنشر بتأريخ 2019/1/15**

ISSN : 1813-6729 http://doi.org/10.31272/JAE.42.2019.119.4

# I- The Preface and Purpose:

Linear discriminant analysis $(\mathrm{L.D.A})$ is a statistical method to find a linear combination of features that separate between two or more groups of objects or events. [9]

That is, the homoscedasticity of the covariance matrices of the groups and that they are with full rank. The $(\mathrm{L.D.A})$ is similar for linear regression analysis $(\mathrm{L.R.A})$ in terms of being also expressing the dependent variable by linear function of vector of independent variables. The difference between the two analyses is that the dependent variable in $(\mathrm{L.D.A})$ is qualitative (nominal) whereas it is quantitative in $(\mathrm{L.R.A})$. It was developed by Sir Ronald Fisher in 1936. [3] Then it was used widely by several researchers like, Lachen Bruch in 1975 and Klecka William in 1980. [7],[6]

The linear discriminant function $(\mathrm{L.D.F})$ is useful for determining whether a set of variables is effective in predicting category membership. [4] There are three famous rules of discrimination: maximum likelihood, Bayes discriminant rule and Fisher Linear discriminant rule. [5] The purpose of this research is to show the importance of Mahalnobis statistic not only in differentiating between two groups but also in evaluating the linear discriminant function $(\mathrm{L.D.F})$. So the problem here is how to do that mathematically (theoretically and practically).

In order to show that, the research is divided into five sections; First is to define $(\mathrm{D}^2)$ by defining the Euclidean and statistical distances. Second to find the relation between $\mathrm{D}^2$ and $\mathrm{F}$ , $\mathrm{T}^2$ – statistics. Third, to prove that F-statistic for separating between two groups in $(\mathrm{L.D.A})$ is the same for $(\mathrm{L.R.A})$ and to use $(\mathrm{D}^2)$ to derive a rule of discrimination between two groups from ($\mathrm{F}$ - statistic). Fourth, to apply $(\mathrm{D}^2)$ and the new rule of discrimination $(\mathrm{R}_{\mathrm{D}^2})$ to separate between two groups and evaluate $(\mathrm{L.D.F})$. whereas the fifth is devoted to conclusions and recommendations.

# II- Euclidean and Statistical Distances:

The Euclidean distance $(\mathrm{d_E})$ between the two points; $\underline{\mathrm{X}}_{(1)} = (\mathrm{X}_{(1)1}, \mathrm{X}_{(1)2}, \dots , \mathrm{X}_{(1)k})^{\mathrm{T}}$ and $\underline{\mathrm{X}}_{(2)} = (\mathrm{X}_{(2)1}, \mathrm{X}_{(2)2}, \dots , \mathrm{X}_{(2)k})^{\mathrm{T}}$ in space with dimensions $(\mathrm{K})$ can be defined as follows:

$$\mathrm{d_E}(\underline{\mathrm{X}}_{(1)}, \underline{\mathrm{X}}_{(2)}) = \sqrt{(\mathrm{X}_{(1)1} - \mathrm{X}_{(2)1})^2 + \cdots + (\mathrm{X}_{(1)k} - \mathrm{X}_{(2)k})^2}$$
$$= \sqrt{(\underline{\mathrm{X}}_{(1)} - \underline{\mathrm{X}}_{(2)})^{\mathrm{T}}(\underline{\mathrm{X}}_{(1)} - \underline{\mathrm{X}}_{(2)})}$$

ISSN : 1813-6729 http://doi.org/10.31272/JAE.42.2019.119.4

But this distance does not take in the account the difference of scales, so, to make them standard, we have to divide
$X_{(1)i} \& X_{(2)i}$ by $S_i$ ; $i = 1, 2, \ldots\ldots, K$ ; Then :

$$\underline{X}^*_{(1)} = \left(\frac{X_{(1)1}}{S_1}, \ldots., \frac{X_{(1)k}}{S_K}\right) \& \underline{X}^*_{(2)} = \left(\frac{X_{(2)1}}{S_1}, \ldots., \frac{X_{(2)k}}{S_K}\right)$$

$$\therefore d\,(X_{(1)}, X_{(2)}) = d_E\left(X^*_{(1)}, X^*_{(2)}\right) = \sqrt{\left(\frac{X_{(1)1} - X_{(2)1}}{S_1}\right)^2 + \cdots + \left(\frac{X_{(1)k} - X_{(2)k}}{S_k}\right)^2}$$

$$= \sqrt{(\underline{X}_{(1)} - \underline{X}_{(2)})^T D^{-1} (\underline{X}_{(1)} - \underline{X}_{(2)})}$$

Where D = diag($S_1^2$ , $S_2^2$ , … , $S_k^2$).

also this distance does not take in account the homoscedasticity and the correlation between $\underline{X}_{(1)} \& \underline{X}_{(2)}$. But with taking them in account and denoting the pooled within covariance matrix by $V_P$ , then the statistical distance will be:

$\therefore d_s\,(\underline{X}_{(1)}, \underline{X}_{(2)}) = \sqrt{(X_{(1)} - X_{(2)})^T V_P^{-1} (X_{(1)} - X_{(2)})}$ which is called Mahalnobis distance. The Mahalnobis distance from centre$\underline{X}_{(1)}$ to centre $\underline{X}_{(2)}$is denoted as following:

$$\therefore d_s\left(\overline{X}_{(1)}, \overline{X}_{(2)}\right) = \sqrt{\left(\overline{X}_{(1)} - \overline{X}_{(2)}\right)^T V_P^{-1} \left(\overline{X}_{(1)} - \overline{X}_{(2)}\right)} = D$$

And the square of this distance is called Mahalnobis statistics ($D^2$) where:

$$D^2 = \left(\overline{X}_{(1)} - \overline{X}_{(2)}\right)^T V_P^{-1} \left(\overline{X}_{(1)} - \overline{X}_{(2)}\right) \ldots\ldots\ldots\ldots\ldots\ldots \text{(1)}$$

Which is very useful for evaluation of (L. D. F) between two groups and discovering the outliers.

### III-The relation between $D^2$ and F , $T^2$:

For univariate normal data (UND) of two independent random samples; $X_1 = (X_{11}, X_{12}, \ldots, X_{1n1})^T \& X_2 = (X_{21}, X_{22}, \ldots, X_{2n2})^T$ with honogenuous variances:

$$t = \frac{(\overline{X}_1 - \overline{X}_2) - (M_1 - M_2)}{V_P\left(\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right)} \quad \text{and under } H_0 : M_1 = M_2$$

$$t = \frac{\overline{X}_1 - \overline{X}_2}{V_P\left(\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right)} = \frac{\frac{\overline{X}_1 - \overline{X}_2}{V_P}}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$ and by squaring the two sides and simplifying

ISSN : 1813-6729 http://doi.org/10.31272/JAE.42.2019.119.4

it:

$$t^2 = \frac{n_1\, n_2}{n_1 + n_2}\left(\frac{\bar{X}_1 - \bar{X}_2}{V_P}\right)^2 \text{ ……………………. (2)}$$

$$\therefore t^2 \sim F\,(1\,, n_1 + n_2 - 2)$$

Now let's have multivariate normal data (MND) of two independent random samples with the same covariance matrix of full rank($V_P$), then the relation (2) becomes:

$$T^2 = \frac{n_1\, n_2}{n_1 + n_2}(\bar{X}_{(1)} - \bar{X}_{(2)})^T V_P^{-1}(\bar{X}_{(1)} - \bar{X}_{(2)}) \text{………… (3)}$$

Where $V_P$ is the pooled variance-covariance matrix for both $X_{(1)}\&X_{(2)}$. $T^2$is hoteling statistic.

But from (1) : $D^2 = \left(\bar{X}_{(1)} - \bar{X}_{(2)}\right)^T V_P^{-1}\left(\bar{X}_{(1)} - \bar{X}_{(2)}\right)$ then:

$$T^2 = \frac{n_1\, n_2}{n_1 + n_2} D^2 \text{………………. (4)}$$

$$\because \frac{n_1 + n_2 - K - 1}{K\,(n_1 + n_{2-2})} T^2 \sim F\,(K\,, n_1 + n_2 - K - 1) \text{………………. (5)}$$

Then $F = \frac{n_1\, n_2\,(n_1 + n_2 - K - 1)}{K\,(n_1 + n_2)(n_1 + n_{2-2})} D^2$ ………………. (6)

or $D^2 = \frac{K\,(n_1 + n_2)(n_1 + n_{2-2})}{n_1\, n_2 (n_1 + n_2 - K - 1)} F$ ………………. (7)

## IV-The Same Fisher Ratio ($F$) in($L.D.A$) and ($L.R.A$) :

Let's have multivariate data for ($\underline{X}$) represents two groups of random observations; $\underline{X}_{(1)}$, $\underline{X}_{(2)}$belonging to two populations distributed normally with means $\underline{M}_{(1)}\&\underline{M}_{(2)}$ and covariance matrices $\sum(1)$ , $\sum(2)$. Then ($L.D.A$) requires the following main assumptions: [1]

1) $P\left(X_{(1)}\right) = P\left(X_{(2)}\right) = P \sim MND\left(\underline{M}_{(1)}, \Sigma_{(1)}\right)\&\left(\underline{M}_{(2)}, \Sigma_{(2)}\right)$ respectively.
2) $\Sigma_{(1)} = \Sigma_{(2)} = \Sigma$ (with full rank) $\Rightarrow$ Homoscedasticity.

Fisher discriminant analysis ($F.D.A$) will be the same as ($L.D.A$) if it satisfies the above assumptions.

Let $\underline{w}$ be the vector of coefficients of ($L.D.F$) between two groups, then $\underline{w}\,.\,\underline{x}$ is distributed normally with means $\underline{w}\,.\,\underline{M}_{(i)}$ and variance $\underline{w}^T\Sigma_{(i)}\underline{w}$ for (i = 1, 2).

The Fisher rule of discrimination (separation) between two groups will be the ratio of the variance between groups to the variance of within groups as follows:

$$\text{Discrimination} = \frac{\sigma_B^2}{\sigma_w^2} = \frac{(\underline{W}\,.\underline{M}_{(1)} - \underline{W}\,.\underline{M}_{(2)})^2}{\underline{W}^T\Sigma_{(1)}\underline{W} + \underline{W}^T\Sigma_{(2)}\underline{W}} = \frac{[\underline{W}(\underline{M}_{(1)} - \underline{M}_{(2)})]}{\underline{W}^T(\Sigma_{(1)} + \Sigma_{(2)})\,\underline{W}} \text{ ……. (8)}$$

ISSN : 1813-6729 http://doi.org/10.31272/JAE.42.2019.119.4

That is, the best discrimination occurs when:

$$\underline{w} \propto (\Sigma_{(1)} + \Sigma_{(2)})^{-1}(\underline{M}_{(1)} - \underline{M}_{(2)})^2$$

And according to the assumption (2) the best discrimination will be:

$$\underline{w} \propto \frac{1}{2}\Sigma^{-1}(\underline{M}_{(1)} - \underline{M}_{(2)})^2$$

With symbols of statistics from random samples the vector will be:

$$\underline{w} \propto \frac{1}{2}V_P^{-1}(\overline{\underline{X}}_{(1)} - \overline{\underline{X}}_{(2)})^2 \Longrightarrow \underline{w}\&\frac{1}{2}D^2 \ldots\ldots\ldots\ldots\ldots\ldots \quad (9)$$

So $(w)$ the vector of coefficients of $(L.D.F)$ depends entirely on Mahalnobis statistic $(D^2)$ to separate between two groups. Now since p.d.f of the two groups is the same from assumption (1). Then: [2]

$$P\left(X_{(1)}\right) = P\left(X_{(2)}\right) = P = \emptyset\left[\frac{-1}{2}\sqrt{\left(\overline{X}_{(1)} - \overline{X}_{(2)}\right)^T V_P^{-1}(X_{(1)} + X_{(2)})}\right] = \emptyset\left(-\frac{1}{2}\sqrt{\mathrm{D}^2}\right) \ldots\ldots \quad (10)$$

which is used to find the probability of real misclassification. But the rule of discrimination in (8) is (F-statistic) itself and it is the same in $(L.D.A)$ and $(L.R.A)$ because it is a ratio of the same two variances $(MSB \,\&\, MSW)$ between groups and within groups. So we can compute (F) from the ANOVA of $(L.R)$ and use it to find$(D^2)$from the relation (7).

The two variances of ratio (F) in $(L.D.A)$ can be written in terms of $D^2$ as follows:

$$SSW = D^2 \Longrightarrow MSW = \frac{D^2}{n_1 + n_2 - K - 1} \ldots\ldots\ldots\ldots\ldots \quad (11)$$

But $F = \frac{MSB}{MSW} \Longrightarrow MSB = MSW.F \; and \; from \; (6 \,\&\, 11)$:

$$MSB = \frac{D^2}{n_1 + n_2 - K - 1} \cdot \frac{n_1 n_2(n_1 + n_2 - k - 1)}{K\,(n_1 + n_2)(n_1 + n_2 - 2)} \cdot D^2$$

$$MSB = \frac{n_1 n_2}{K\,(n_1 + n_2)(n_1 + n_2 - 2)} \cdot (D^2)^2 \ldots\ldots\ldots\ldots\ldots \quad (12)$$

Then the discriminant rule $(F)$ can be written as $(R_{D^2})$

$$\therefore\; R_{D^2} = \frac{MSB}{MSW} = \frac{D^2}{n_1 + n_2 - K - 1} \div \frac{n_1 n_2}{K\,(n_1 + n_2)(n_1 + n_2 - 2)}$$

$$R_{D^2} = \frac{n_1 n_2\,(n_1 + n_2 - K - 1)}{K\,(n_1 + n_2)(n_1 + n_2 - 2)} \cdot D^2 \sim F\,(K, n_1 + n_2 - K - 1) \ldots\ldots \quad (13)$$

So the discrimination between two groups depends entirely on Mahalnobis statistic $(D^2)$ . In addition, the vector of discriminant coefficients in (9) which is proportional to the Mahalnobis statistic, can

ISSN : 1813-6729 http://doi.org/10.31272/JAE.42.2019.119.4

be found using $(D^2)$ itself and the multiple linear regression coefficients as follows: [8] $\widehat{W}_i = \hat{B}_i\left(\frac{1}{C}\right)$ ; $i = 1, \ldots, K$.......... (14)

Where $\frac{1}{C} = \frac{n_1 + n_2}{n_1 n_2}(n_1 + n_2 - 2) + D^2$.......... (15)

It is worth mentioning that the value of $D^2$ and the discriminant coefficients can be found also (directly) from multivariate data by finding the pooled variance-covariance matrix (within groups) as following:

$\underline{\widehat{W}} = V_P^{-1}\left(\overline{X}_{(1)} - \overline{X}_{(2)}\right) \& D^2 = \widehat{W}^T\left(\overline{X}_{(1)} - \overline{X}_{(2)}\right)$

## V- The Application:

To show the importance of Mahalnobis statistic $(D^2)$ in discrimination between two groups by application, the researcher has gathered data randomly on some variables which are (age, weight and blood pressure) from (40) patients at (heart disease unit) in Medical City within four days. He has found that (16) of them suffers from coronary heart disease (CHD) but the rest don't. Patients have been divided into two groups (with CHD & without) and used SPSS is used on the data and got the following:

1) The assumptions of $(L.D.A)$ are satisfied:
   a- Normal distribution of the vector of random variables ($\underline{X}$) as in the table (1): see the tables in the appendix.
   b- Homogenuous variance – covariance matrices as in table (2).
   c- No significant impact of multicollinearity, where all $VIF < 5$ as in table (3).
   d- No outliers according to the values of Mahalnobis distance as compared to $X^2(0.005,2) = 10.597$
2) The estimates of parameters of $(L.R.F)$ except Bo and the value of F-statistic are as follows:
   $\underline{\widehat{B}}^T = [0.006, 0.018, 0.014]$ and $F = 15.893$
3) By applying (7) ; $D^2 = 5.24248$
4) By using $D^2$ and the relations (11 , 12 , 13) , ANOVA of $(L.D.A)$ is created (see table 4) where; $MSW = 0.1456245$ and $MSB = 2.3144105$, then the new rule of discrimination is :

$R_{D^2} = 15.893$, to be compared with $F(3, 36, 0.01) = 4.51$ . It is obvious that the vector $\underline{X}$ is significant and reliable for classifying the patients.

5) By applying (14 , 15) : $\frac{1}{C} = 12.0277$

$$\underline{\widehat{W}}^T = [0.072, 0.216, 0.168]$$

6) The $(L.D.F)$ is: $\hat{Z} = 0.072X_1 + 0.216X_2 + 0.168X_3$

ISSN : 1813-6729 http://doi.org/10.31272/JAE.42.2019.119.4

7) The evaluation of $(\mathrm{L.D.F})$:

a) The rate of apparent error in classification equals 0.10 [see the table (5)].

b) The rate (probability) of real error in classification equals:

$$\mathrm{P} = \emptyset\left(-\frac{1}{2}\sqrt{\mathrm{D}^2}\right) = \emptyset\left(-\frac{1}{2}\sqrt{5.242}\right) = \emptyset(-1.1448)$$
$$= 1 - 0.87285 = 0.127 \cong 0.13$$

The probability of misclassification is little, that is, the $(\mathrm{L.D.F})$ for patients is significant and reliable.

## VI- Conclusions & Recommendations :

a-Conclusions : we can conclude the following:

1- The Mahalnobis statistic $(\mathrm{D}^2)$ is the base to create a significant rule of discrimination between two groups.

2- The probability of real misclassification for any $(\mathrm{L.D.F})$ can be found directly as a probability function of $\mathrm{D}^2$ .

3- The vector of discriminant coefficients can be easily found from the regression coefficients with the help of Mahalnobis statistic.

4- Any vector of random variables distributed normally can be tested easily and quickly whether it is significant for discrimination or not by the use of Mahalnobis statistic.

b- Recommendation: I recommend using Mahalnobis statistic for the discrimination between two groups and evaluation of any $(\mathrm{L.D.F})$.

## References :

1) Bökeoğlu COKLUK, (2008), "Discriminant Function Analysis ; Concept and Application"Eğitimaraştirmalaridergisi, (33), 73-92.

2) Bloch, B., W., and C.J. Huang, (1974), "Multivariate Statistical Methods for Business and Economics", Prentice-Hall, Inc, New Jersy.

3) Cohen et al., (2005), "Applied multiple Regression / Correlation Analysis for The Behavioural Sciences", 3rd ed., Taylor & Francis group.

4) Green, S.B., Salkend, N. J. & A Key, T. M., (2008), "Using SPSS for Windows and Macintosh, Analyzing and Understanding Data", New Jersy, Prentice Hall.

5) Hardle, W. Simar, L., (2007), "Applied Multivariate Statistical Analysis", Springer, Berlin Heidelberg, pp. 289-303 .

6) Klecka, William R., (1980), "Discriminant Analysis ; Quantitative Applications in The Social Sciences", series No.19 , sage publications.

ISSN : 1813-6729 http://doi.org/10.31272/JAE.42.2019.119.4

7) Lachenbruch, P. A., (1975), "Discriminant Analysis", NY ; Hafner.
8) AL-Rawi, Dr. Khashia, (1987), "Introduction to Regression Analysis", ALmosul University, Book House for publication.
9) Wikipedia ; Free encyclopedia, 2016 .

## The appendix

Table (1)

Kolmogorov –Smirnov Test of normality

| | $X_1$ | $X_2$ | $X_3$ |
|---|---|---|---|
| Asymp. Sig. (2 – tailed) | 0.506 | 0.403 | 0.343 |

Table (2)

Box's M-Test for homogeneity

| | |
|---|---|
| Box's M | 10.559 |
| Sig. | 0.143 |

Table (3)

The impact of multicollinearity

| Variables | VIF |
|---|---|
| $X_1$ | 1.077 |
| $X_2$ | 1.249 |
| $X_3$ | 1.166 |

Table (4)

ANOVA of $(L.D.A)$

| S.O.V. | d.f | S.S | M.S | F |
|---|---|---|---|---|
| Between X's | 3 | 6.943231382 | 2.314410461 | 15.893 |
| Within X's | 36 | 5.242482639 | 0.1456624517 | |
| Total | 39 | 12.185714021 | | |

Table (5)

The classification of patients

| Group : G | Related to $G_1$ | Related to $G_2$ | Sum |
|---|---|---|---|
| 1 | 16 | 0 | 16 |
| 2 | 4 | 20 | 24 |
| % | 100<br>16.7 | 0<br>83.3 | 100<br>100 |

90% of original grouped cases are correctly classified.

……………………………………………………………………

……………………………………………………………