International Mathematical Forum, Vol. 4, 2009, no. 8, 351-357

Discriminant Analysis: a case study of a war data set

M. Chalikias[1], G. Kaimakamis, M. Adam and N. Karadimas

Hellenic Army Academy, 16673 Vari Attika, Greece.

Abstract

Mathematical models are applied in war theories as these of Richarson and Lanchester. In this paper discriminant analysis is used for the most famous battles of the Second World War. The linear function of Fisher classifies the opposite sides in two groups (the one of the winners and the one of the losers). That function can be used in order to predict the winner of a battle and to evaluate the difficulty of a win.

Keywords : Discriminant Analysis; Fisher rule; Data Analysis
Mathematics Subject Classification : 62-07

1 Introduction

Historically, the first mathematical expressions about the connections between force sizes and loss were developed from J.V. Chase in [3]. Later F.W. Lanchester [5] showed the importance of concentration of troops in modern combat. Lanchester constructed mathematical formulas called "Lanchester laws", which were given by a set of ordinary differential equations. These equations can be treated as models of attrition in modern warfare. For example, Lanchester laws include the idea that, the attrition rate of one side is proportional to the opposing side's size in the case of directed fires [7]. Many researchers tried to generalize Lanchester's laws by using tools from the field of Stochastic and Partial Differential Equations [1, 8, 9]. In [4] the author gives an introduction in the theory of war's stochastic processes.

The present paper deals with the construction of a linear function with several variables which can predict the result of a battle under given circumstances. For this Discriminant Analysis has been used and especially the Discriminant Rule of Fisher [2, 6].

[1]Corresponding author, e-mail : *mchalikias@ucg.gr*

2 Data and Methods

The data sets concern the Second World War battles. The results of these battles are known and the winner was either the allays's or the axis's forces. For each battle we use the data, which are presented in the following tables:

Table 1. Loser group

Forces	Battles	Troops	Tanks	Losses
Axis's	Stalingrad	1011000	675	850000
Axis's	Kursk	800000	2700	50000
Allays's	Ardennes, Belgium	500000	500	84834
Axis's	Monte Cassino, Italy	80000		20000
Allays's	El Alamein, Egypt	96000	585	17000
Axis's	El Alamein, Egypt II**	116000	547	37956
Axis's	Normandy, France	380000		9000
Allays's	France, Low Countries	4050000	2445	158830

Table 2. Winners group

Forces	Battles	Troops	Tanks	Losses
Allays's	Stalingrad	1103000	1463	750000
Allays's	Kursk	1300000	3600	180000
Axis's	Ardennes, Belgium	830000	424	85913
Allays's	Monte Cassino, Italy	105000		54000
Axis's	El Alamein, Egypt	150000	1114	13250
Allays's	El Alamein, Egypt II**	220000	1029	13900
Allays's	Normandy, France	1000000		10264
Axis's	France, Low Countries	2862000	3384	2260000

Our purpose is to create a linear discriminant function (rule) which can classify each of these sides (allays and axis) to the winners. Furthermore, this function will give the possibility to predict the winner of a battle, if specific data for the fighting sides such as the number of troops and tanks are known. We construct the following variables in order to make the data of these battles comparable and suitable for the discriminant analysis :

$$\text{Troops} = \frac{\text{number of soldiers from } A \text{ army}}{\text{number of soldiers from } B \text{ army}}$$

$$\text{Tanks} = \frac{\text{number of tanks that } A \text{ army had}}{\text{number of tanks that } B \text{ army had}}$$

$$\text{Human Losses} = \frac{\text{number of soldiers who died or were injured/captured}}{\text{number of soldiers at the beginning of the battle}}$$

A discriminant function presupposes the existence of two groups to be separated, in our case we have the group of the winners and the one of losers. Each group takes a score from the discriminant function, let's say $u_{ij},\ i = 1, 2$ and $j = 1, 2, \ldots, n_i$, where n_i is the number of the observations for every variable. The main goal is to find a function which maximizes the difference of the mean $\overline{u}_1 - \overline{u}_2$. We can use as a measure of distance, the quantity

$$D = \frac{|\overline{u}_1 - \overline{u}_2|}{S_U}$$

where

$$S_U = \frac{\sum_{j=1}^{n_1}(u_{1j} - \overline{u}_1)^2 + \sum_{j=1}^{n_2}(u_{2j} - \overline{u}_2)^2}{n_1 + n_2 - 2}.$$

We want to maximize D or equivalently maximize D^2. Discriminant Fisher function is a linear function of the form

$$u_{ij} = Lx_{ij},$$

where $x_{ij},\ (i = 1, 2,\quad j = 1, 2, \ldots, n_i)$ is the vector with the values of the variables which are used in the model, (in our case the variables are Troops, Tanks, Human Losses) and L is the vector of the coefficients.

Let L^t denote the transpose of the vector L. Thus, combining the above relationships we have to maximize the quantity

$$D^2 = \frac{\left(L^t(\overline{x}_1 - \overline{x}_2)\right)^2}{L^t S_p L}. \tag{1}$$

We remind the known Cauchy-Schwartz inequality, where in Euclidean vector space $\mathbb{R}^p$ using the standard inner product, for every vector $a, b \in \mathbb{R}^p$ holds

$$\left(a^t b\right)^2 \leq \left(a^t a\right)\left(b^t b\right). \tag{2}$$

If we suppose that the covariance matrix is positive definite, substituting in (2) $a = S_p^{1/2} L$ and $b = S_p^{-1/2}(\overline{x}_1 - \overline{x}_2)$, we have

$$\begin{aligned} \left(L^t(\overline{x}_1 - \overline{x}_2)\right)^2 &\leq \left(L^t S_p^{1/2} S_p^{1/2} L\right)\left((\overline{x}_1 - \overline{x}_2)^t S_p^{-1/2} S_p^{-1/2}(\overline{x}_1 - \overline{x}_2)\right) \Leftrightarrow \\ \left(L^t(\overline{x}_1 - \overline{x}_2)\right)^2 &\leq \left(L^t S_p L\right)\left((\overline{x}_1 - \overline{x}_2)^t S_p^{-1}(\overline{x}_1 - \overline{x}_2)\right). \end{aligned}$$

Due to (1) the last inequality can be written

$$D^2 \leq \left((\overline{x}_1 - \overline{x}_2)^t S_p^{-1}(\overline{x}_1 - \overline{x}_2)\right).$$

If $L = cS_p^{-1}(\overline{x}_1 - \overline{x}_2)$, with $c > 0$, we have

$$D^2 = (\overline{x}_1 - \overline{x}_2)^t S_p^{-1}(\overline{x}_1 - \overline{x}_2),$$

which is the maximum distance. In this case we get the best possible segregation. The discriminant function is complete, if we define as critical value the following quantity:

$$m = \frac{\overline{u}_1 - \overline{u}_2}{2} = \frac{L^t(\overline{x}_1 - \overline{x}_2)}{2}$$

Then the discriminant rule states:

"if $L^t x \geq m \Leftrightarrow L^t x - m \geq 0$, then classified in the first group"

In this paper the score of the discriminant function is calculated by the usage of the program SPSS16.

3 Descriptive Statistics

First of all descriptive statistics of each group are available in the following table, where denoted by 0 correspond to winner's group and 1 correspond to loser's group.

Table 3. Group Statistics

grouping		Mean	Std. Deviation	Valid N Unweighted	(listwise) Weighted
0	Troops	1,42362000	0,439099142	6	6,000
	Tanks	1,58637200	0,484917668	6	6,000
	Human Losses	0,31051794	0,331379538	6	6,000
1	Troops	0,78612541	0,335822711	6	6,000
	Tanks	0,69497708	0,263974502	6	6,000
	Human Losses	0,26940454	0,298008379	6	6,000
Total	Troops	1,10487200	0,499738219	12	12,000
	Tanks	1,14067500	0,596039531	12	12,000
	Human Losses	0,28996124	0,301236527	12	12,000

It is remarkable that the means of Human Losses in each group are very close.

For the selection of the most parsimonious model stepwise method has been used. The final model includes only the variables Troops and Tanks. ANOVA table indicates that the discrimination of the two variables from the Fisher's function is very satisfactory (statistically significant in the level of 1%).

Table 4. Wilks'λ for the two variables

						Exact F			
Step	Variables	λ	df1	df2	df3	Statistic	df1	df2	Sig.
1	Troops	0,390	1	1	10	15,640	1	10,000	0,003
2	Tanks	0,279	2	1	10	11,605	2	9,000	0,003

Moreover the values of Wilks'λ for the two variables give the percentage of variance which can't be explained with the model (the Fisher's function). We should consider as satisfactory values the values less than 0,5.

The assumption for the equality of variances of the two variables can't be rejected as we observe in the following Box's Matrix (see in Table 5).

Table 5. Box's Matrix[3]

Box's M	2,109
Approx.	0,551
df1	3
df2	18000,000
Sig.	0,648

Wilks'λ of the model is under $0,5$ and gives a satisfactory p-value (less than $0,05$).

Table 6. Wilks'λ for the model

Test of function(s)	Wilks'λ	Chi-square	df	Sig.
1	0,279	11,476	2	0,003

The discriminant Fisher's function is given in Table 7. According to that each of opposite sides (allays's and axis's forces) group is classified in group 0 (group of winners), if the function $w_1 = -17,013 + 10,31\text{Troops} + 11,31\text{Tanks}$ is greater than the score of $w_2 = -4,64 + 5,59\text{Troops} + 5,05\text{Tanks}$. Equivalently for positive values of the function $w = -17,013 - (-4,464) + (10,31 - 5,59)\text{Troops} + (11,31 - 5,05)\text{Tanks}$ equals to $w = 12,549 + 4.72\text{Troops} + 6.26\text{Tanks}$ the group is classified in winners group else it is classified if losers group.

Table 7. Fisher's linear discriminant functions

	grouping	
	0	1
Troops	10,314	5,590
Tanks	11,319	5,053
(constant)	$-17,013$	$-4,646$

Next table summarizes the results of the Fisher's function, where $91,7\%$ of the grouped cases are correctly classified.

Table 8. Classification results[4]

grouping			Predicted Group 0	1	Total
Original	Count	0	5	1	6
		1	0	6	6
	%	0	83,3	16,7	100,0
		1	0	100,0	100,0

[3]Tests null hypothesis of equal population covariance matrices.

[4]$91,7\%$ of original grouped cases correctly classified.

From the Casewise Statistic table (see Table 9), we have that misclassified observations is the axis's forces in France battle, which was classified in the group of losers. Moreover, from that table we conclude the discriminant scores of the function w.

Table 9. Casewise Statistics

Case number		Actual Group	Predicted Group	Discriminant Scores
Original	1	1	1	$-1,755$
	2	1	1	$-1,624$
	3	1	1	$-0,727$
	5	1	1	$-2,064$
	6	1	1	$-2,232$
	8	1	1	$-0,394$
	9	0	0	2,172
	10	0	0	1,250
	11	0	0	0,269
	13	0	0	2,369
	14	0	0	2,858
	16	0	1[5]	$-0,112$

Finally the results above are verified from the histograms for each group in the following figure.

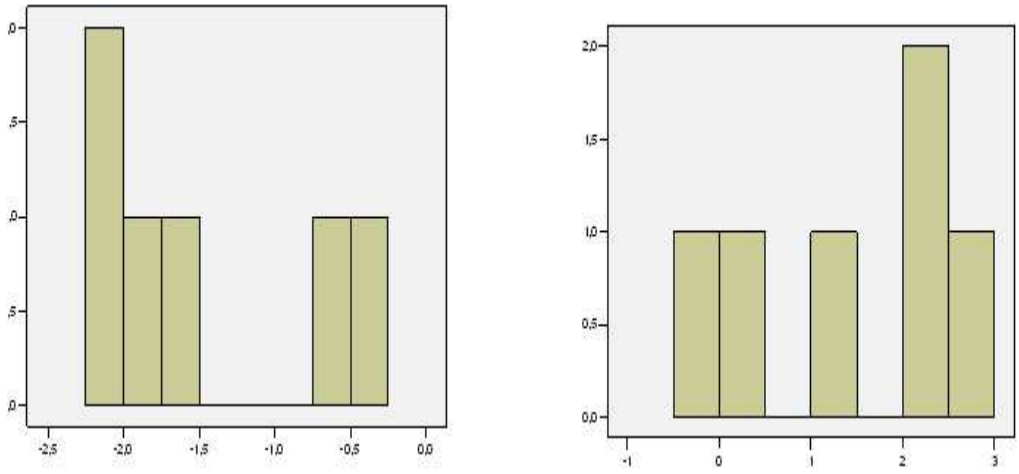

Figure 1: The left histogram corresponds to the winner's group and the right to the loser's group.

[5]Misclassified case.

4 Discussion

Fisher's function classifies the data in a satisfactory level $(91,7\%)$. That fact gives the possibility to predict the winner of a battle. Finally we observe that axis's forces in France battle is the only misclassified observation. This proves that the win in France form axis's forces was earned in unfavorable conditions relative to the opposite side.

References

[1] C.J. Ancher and A.V. Gafarian, Modern Combat models. A critique of their Foundations, *Topics in Operational Research*, ORSA, 1992.

[2] L. Breiman, J. Friedman, R. Olshen and C. Stone, Classification and Regression Trees, *Wodsworth International Group*, 1984.

[3] J.V. Chase, A Mathematics Investigation of the effect of Superiority in Combats Upon the sea, 1902 reprinted by B.A. Fiske, The Navy as a fighting Machine, Annapolis, *US Naval Institute Press*, 1988.

[4] N.J. Daras, Non Lanchesterian Stochastic processes in war analysis Part I : stochastic description of Combat's losses, *Journal of Interdisciplinary Mathematics*, **10**, no.5, (2007), 637-680.

[5] F.W. Lanchester, Aircraft in Warfare *Engineering*, **98** (1914), 422-423; reprinted in *The World of Mathematics*, **IV** (1956), 2138-2148, edited by Newman, Simon and Schuster.

[6] G.J. Mclachlan, *Discriminant analysis and statistical pattern recognition*, Willey, New York, 1992.

[7] R. Mquie, Military history and Mathematical Analysis, *Military Review*, **50**, no.5, (1970), 8-17.

[8] L.F. Richardson, Mathematics of war and foreign politics, *The World of Mathematics*, **4**, Edited Neuman Simon and Schuster, 1240-1253.

[9] J.G. Taylor, Lanchester Models of Warfare 2 Vols, *Military Applications Section of the Operations Research Society of America*, 1983.

Received: June 30, 2008