PETRESCU Mary-Violeta

METODE DE APLICARE A ANALIZEI DISCRIMINANTE PENTRU DOUĂ POPULAȚII

Craiova
Ianuarie, 2025

Copyright © 2025, Autor
Tehnoredactare computerizată: PETRESCU Mary-Violeta
Corectura aparţine autorului

Responsabilitatea pentru conţinutul lucrării revine în exclusivitate autorului
Studiu științific

Publicat: ianuarie 2025
http://www.edituraelse.ro/ped/ped.htm

METODE DE APLICARE A ANALIZEI DISCRIMINANTE PENTRU DOUĂ POPULAȚII

Analiza discriminantă este o metodă care face parte din grupa metodelor explicative de analiză a datelor. Aceasta metodă utilizează o variabilă de explicat Y și mai multe variabile explicative cantitative sau binare notate cu $(X_1, X_2, ..., X_p)$.

Principalele obiective ale analizei discriminante sunt:

- determinarea variabilelor explicative care contribuie cel mai mult la diferenţierea claselor definite de variabila de explicat pentru a identifica ponderea influenţei variabilelor explicative în variaţia variabilei de explicat prin construirea de combinaţii liniare în scopul alegerii celei mai bune;
- construirea unui spaţiu discriminant în funcţie de tipul analizei discriminante, adică: pentru analiza discriminantă liniară unde variabilă de explicat are două stări se determină o axă discriminantă Z care explică apartenenţa unei unităţi la o clasă sau alta, iar pentru analiza discriminantă multiplă se determină mai multe combinaţii liniare independente de variabile explicative numite (axe discriminante) fiind necesară analiza spaţiului determinat de acele axe care separă cel mai bine unităţile studiate, în clasele determinate de stările variabilei de explicat Y;
- repartizarea la o clasă existentă a unităţilor care nu au fost clasate, cunoscându-se nivelul variabilelor explicative şi utilizându-se funcţiile discriminant care vor fi repartizate noilor unităţi în clasele deja formate.

Metoda se aplica unei populații de indivizi caracterizate prin variabile continue sau categoriale și ale cărei componente sunt împărțite în grupuri. Scopul principal al analizei discriminante este clasificarea uneia sau mai multor observații noi în aceste grupuri deja precizate.

În continuare, va fi enunțată problema analizei discriminante.

Fiind dată o variabilă de explicat Y cu *k* grupe (stări) şi*p* variabile explicative $(X_1, X_2, ..., X_p)$, trebuie găsită una sau mai multe combinaţii liniare de variabile explicative de forma:

$$Z = \sum_{i=1}^{p} a_i X_i \qquad (1)$$

Combinația liniară din relația (1) diferenţiază cel mai bine grupele formate prin raportare la stările variabilei de explicat Y.

Este cunoscut faptul că matricea de varianţă-covarinţă totală notată cu T, se descompune în două componente:

- matricea de varianţă-covarianţă între grupe notată cu B;
- matricea de varianţă-covarianţă din interiorul grupelor notată cu W, determinată ca o sumă de *k*matrici de varianţă-covarianţă din cadrul grupei.

$$T=B+W \qquad (2)$$

În continuare, vom efectua o analiză a varianţei, utilizând ecuația de descompunere a varianţei totale pentru o variabilă X_i, în scopul determinării puterii de discriminare a fiecărei variabile.Varianţa totală este egală cu suma varianţei dintre grupe şi varianţa din grupe.

Ecuația de descompunere a varianţei totale pentru o variabilă X_i are următoarea formă:

$$\sum_{h=1}^{k}\sum_{j=1}^{N_h}\left(X_{ijh} - \overline{X_i}\right)^2 = \sum_{h=1}^{n} N_h\left(\overline{X_{ih}} - \overline{X_i}\right)^2 + \sum_{h=1}^{k}\sum_{j=1}^{N_h}\left(X_{ijh} - \overline{X_{ih}}\right)^2 \qquad (3)$$

Pentru măsurarea puterii de discriminare a variabilei X_i, se utilizează urmîtorul raport de determinaţie:

$$r^2(X_i, Y) = \frac{\text{Variatia dintre grupe}}{\text{Variatia totala}} \qquad (4)$$

Raportuldinrelația (4) cu cât tinde mai multcătrevaloarea1 cu atât puterea de discriminare a variabilei X_i este mai mare.

Vomutilizavariabila Fisher notată cu F care permite aflarea variabilelor care sunt semnificativ discriminante, având un nivel de semnificaţie $P = prob\left(F(k-1; N-k) \geq F\right)$.

Variabila Fisher F este de forma:

$$F = \frac{\text{Variatia dintre grupe} / k-1}{\text{Variatia din grupe} / N-k} \qquad (5)$$

Problema analizei discriminante constă în căutarea uneia sau a mai multor direcţii în care varianţa totală T se descompune în cele două componente ale sale, efectuându-se operația de minimizare a lui W şi operația de maximizarea lui B.

Axele discriminante Z sunt considerate ca fiind vectorii proprii ai matricii $\left(T^{-1}B\right)$, iar numărul axelor discriminante Z este egal cu min(p;k-1).

Variabilei discriminante (Z_1) îi este asociată cea mai mare valoare proprie, și variabilei discriminante (Z_2) îi corespunde cea de-a doua valoare proprie, fiind necorelată cu prima variabilă $cor(Z_1, Z_2) = 0$.

Variabilei discriminante (Z_1) îi va corespunde cel mai mare raport de corelaţie, iar raportul de corelaţie dintre (Z_2)şi variabila de explicat Y , notat cu $R(Z_2, Y)$ este mai mic decât în cazul primei variabile discriminante, ş.a.m.d.

Prin urmare putem afirma că, variabilele discriminante nu sunt corelate între ele.

În scopul determinării numărului de variabile discriminante ce trebuie luate în considerare se utilizează testul Wilks care determină care sunt ultimele q rapoarte de corelaţie nule. Testul statistic Wilks are următoarea formă:

$$\chi_q = \prod_{m=k-q+1}^{k}\left[1 - R_m^2\left(Z_m, Y\right)\right] \qquad (6)$$

În cazul în care χ are valori mici, atunci ipoteza este respinsă.

Bartlett(1954) şiRao (1973) au efectuat diferite studii cu privire la legile de aproximare a distribuţiei χ_q pentru calcularea nivelului de semnificaţie.

În studiul său, Bartlett a considerat că forma $\chi^2 = -[n-1-(p+q)/2]\ln\lambda_q$ (7)

se poate aproxima cu o lege χ^2 care are *q(k-1)* grade de libertate, în special atunci când numărul de clase *k* este mai mare decât 2 sau 3.

Prin urmare, putem spune că testul lui Wilks măsoară puterea de discriminare totală a axelor noi. Cu cât valoarea lui χ este mai mică cu atât creşte puterea de discriminare a axelor, iar axele de discriminare Z_m care au o putere de discriminare mică nu se vor lua în calcul.

Dacă dorim să interpretăm statistic axele discriminante vom studia legătura dintre axele discriminante şi variabilele explicative cu ajutorul coeficienţilorfuncţiilor şi coeficienţilor de corelaţie dintre axa discriminantă şi fiecare variabilă explicativă.

Pe baza unei reguli de decizie variabilele discriminate sunt utilizate pentru repartizarea unităţilor în grup.

În cazul în care există numai două axe atunci regula de decizie se poate realiza foarte uşor, dar în cazul când sunt mai multe axe acest lucru nu se poate realiza decât dacă se calculează probabilităţile de apartenenţă la grupe diferite, după cum urmează:

$$p_h(x) = Prob(Y = h / X_1 = x_1,..., X_p = x_p) \qquad (8)$$

În cazul în care vectorul X al variabilelor explicative $(X_1, X_2,..., X_p)$ urmează o lege multi-normală $N(\mu_h, \sigma)$probabilităţile pentru fiecare grupă le vom determina după cum urmează:

$$p_h(x) = \frac{exp\, g_h(x)}{\sum_{h=1}^{k} exp\, g_h(x)} \qquad (9)$$

Funcţiile discriminante sunt variabilele $g_h(x)$ fiind estimate prin $\hat{g}_h(x)$, astfel:

$$\hat{g}_h(x) = -\frac{1}{2}\overline{X_h'}S^{-1}\overline{X_h} + \overline{X_h}S^{-1}X' \ , \ \overline{X_h} = (\overline{X_1}, \overline{X_2},..., \overline{X_p}) \quad (10)$$

unde S reprezintă matricea de varianţăşicovarianţă din grupe.

Atunci, putem estima probabilitatea $p_h(x)$, astfel:

$$\hat{p}_h(x) = \frac{exp\,\hat{g}_h(x)}{\sum_{h=1}^{k} exp\,\hat{g}_h(x)} \qquad (11)$$

Conform relației (11) putem spune că fiecare unitate va putea fi repartizată în grupa în care se înregistrează probabilitatea cea mai mare, iar prin utilizarea regulii de decizie optimală stabilite noile unităţi se pot repartiza în grupele deja formate în funcție de nivelul înregistrat de variabilele explicative.

Valorile înregistrate de funcţiile discriminante estimate se pot calcula pentru fiecare nouă unitate ce trebuie repartizată într-o grupă, adică $\hat{g}_h(x)$.

Printre metodele de aplicare a analizei discriminante pentru două populații regăsim metoda de separare şi metoda de clasificare.În continuare vom prezenta în detaliu cele două metode.

- ***Metoda de separare***

Să presupunem că avem două populații. Fie $X_1, X_2, \ldots, X_{n_1}$ cele n_1 observații dinpopulația 1 şi fie $X_{n_1+1}, X_{n_1+2}, \ldots, X_{n_1+n_2}$ cele n_2 observațiidinpopulația 2.

De notat că $X_1, X_2, \ldots, X_{n_1}, X_{n_1+1}, X_{n_1+2}, \ldots, X_{n_1+n_2}$ sunt vectori $p \times 1$–dimensionali.

Metoda discriminantă a lui Fisher constă în proiectarea acestor vectori $p \times 1$ –dimensionali din spaţiul $\Re^{p+1}$ în spaţiul $\Re$ prin intermediul unei funcţii liniare $l(X) = a^t X$ în încercarea de a separa cele două populaţii cât mai mult posibil, unde a este un vector $p \times 1-$ dimensional. În continuare, va fi prezentată metoda discriminantă a lui Fisher după cum urmează:

Să se găsească vectorul $\hat{a}$ ce maximizează funcţia de separaţie $|S(a)|$,

$$S(a) = \frac{\bar{Y}_1 - \bar{Y}_2}{S_Y} \qquad (12)$$

unde $\overline{Y}_1 = \frac{\sum_{i=1}^{n_1} Y_i}{n_1}, \overline{Y}_2 = \frac{\sum_{i=n_1+1}^{n_1+n_2} Y_i}{n_2}, S_Y^2 = \frac{\sum_{i=1}^{n_1}(Y_i - \overline{Y}_1)^2 + \sum_{i=n_1+1}^{n_1+n_2}(Y_i - \overline{Y}_2)^2}{n_1 + n_2 - 2}$, (13)

și $Y_i = a^t X_i, i = 1,2,\ldots,n_1 + n_2$ (14)

Reprezentarea intuitivă a metodei discriminante a lui Fisher, în măsura în care se poate de a găsi $\hat{a}$, este ilustrată grafic în figura de mai jos (a se vedea Figura 1.).

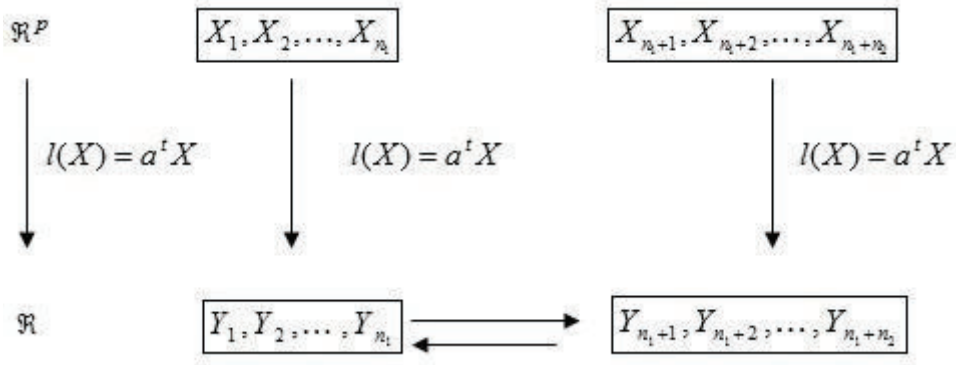


Figura 1. Reprezentarea intuitivă a metodei discriminante a lui Fisher

Intuitiv, $S(a) = \frac{\overline{Y}_1 - \overline{Y}_2}{S_Y}$ măsoară diferența dintre transformate, adică $\overline{Y}_1 - \overline{Y}_2$ în raport cu deviația standard a eșantionului S_Y. Dacă observațiile transformate $Y_1, Y_2, \ldots, Y_{n_1}$ și $Y_{n_1+1}, Y_{n_1+2}, \ldots, Y_{n_1+n_2}$ sunt complet separate, $\left|\overline{Y}_1 - \overline{Y}_2\right|$ trebuie să fie mare ca variația aleatorie a datelor transformate reflectate de S_Y este de asemenea luată în considerare.

Rezultat important:

Vectorul maximizează separarea $|S(a)| = \left| \frac{\overline{Y}_1 - \overline{Y}_2}{S_Y} \right|$ care este de forma:

$$S_{cumulat}^{-1}\left(\overline{X}_1 - \overline{X}_2\right) \tag{15}$$

unde $$S_{cumulat} = \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2}{n_1 + n_2 - 2}, \tag{16}$$

$$S_1 = \frac{\sum_{i=1}^{n_1} (X_i - \overline{X}_1)(X_i - \overline{X}_1)^t}{n_1 - 1} \tag{17}$$

$$S_2 = \frac{\sum_{i=n_1+1}^{n_1+n_2} (X_i - \overline{X}_2)(X_i - \overline{X}_2)^t}{n_2 - 1} \tag{18}$$

și unde $\overline{X}_1 = \frac{\sum_{i=1}^{n_1} X_i}{n_1}$ și $\overline{X}_2 = \frac{\sum_{i=n_1+1}^{n_1+n_2} X_i}{n_2}$ (19)

Justificare:

$$\overline{Y}_1 = \frac{\sum_{i=1}^{n_1} Y_i}{n_1} = \frac{\sum_{i=1}^{n_1} a^t X_i}{n_1} = a^t \left(\frac{\sum_{i=1}^{n_1} X_i}{n_1} \right) = a^t \overline{X}_1 \tag{20}$$

În mod similar, $\overline{Y}_2 = a^t \overline{X}_2$

De asemenea, $\sum_{i=1}^{n_1}(Y_i - \overline{Y}_1)^2 = \sum_{i=1}^{n_1}\left(a^t X_i - a^t \overline{X}_1\right)^2 = \sum_{i=1}^{n_1}\left(a^t X_i - a^t \overline{X}_1\right)\left(a^t X_i - a^t \overline{X}_1\right)^t$

$$= \sum_{i=1}^{n_1} a^t \left(X_i - \overline{X}_1\right)\left(X_i - \overline{X}_1\right)^t a = a^t \left[\sum_{i=1}^{n_1}\left(X_i - \overline{X}_1\right)\left(X_i - \overline{X}_1\right)^t\right] a$$

.

În mod similar, $\sum_{i=n_1+1}^{n_1+n_2}\left(Y_i - \overline{Y}_2\right)^2 = a^t \left[\sum_{i=n_1+1}^{n_1+n_2}\left(X_i - \overline{X}_2\right)\left(X_i - \overline{X}_2\right)^t\right] a$

Astfel, $S_Y^2 = \dfrac{\sum_{i=1}^{n_1}\left(Y_i - \overline{Y}_1\right)^2 + \sum_{i=n_1+1}^{n_1+n_2}\left(Y_i - \overline{Y}_2\right)^2}{n_1 + n_2 - 2}$

$$= \frac{a^t \left[\sum_{i=1}^{n_1}\left(X_i - \overline{X}_1\right)\left(X_i - \overline{X}_1\right)^t\right] a + a^t \left[\sum_{i=n_1+1}^{n_1+n_2}\left(X_i - \overline{X}_2\right)\left(X_i - \overline{X}_2\right)^t\right] a}{n_1 + n_2 - 2}$$

$$= a^t \left[\frac{\sum_{i=1}^{n_1}\left(X_i - \overline{X}_1\right)\left(X_i - \overline{X}_1\right)^t + \sum_{i=n_1+1}^{n_1+n_2}\left(X_i - \overline{X}_2\right)\left(X_i - \overline{X}_2\right)^t}{n_1 + n_2 - 2}\right] a$$

$$= a^t \left[\frac{(n_1 - 1)S_1 + (n_2 - 1)S_2}{n_1 + n_2 - 2}\right] a = a^t S_{cumulat}\, a$$

Astfel, $S(a) = \dfrac{\overline{Y}_1 - \overline{Y}_2}{S_Y} = \dfrac{a^t\left(\overline{X}_1 - \overline{X}_2\right)}{\sqrt{a^t S_{cumulat} a}}\, S(a)$ (21)

$\hat{a}$ pot fi găsite prin rezolvarea ecuației bazat pe prima derivată $S(a)$,

$$\frac{\partial S(a)}{\partial a} = \frac{\left(\overline{X}_1 - \overline{X}_2\right)}{\sqrt{a^t S_{cumulat} a}} - \frac{1}{2} a^t\left(\overline{X}_1 - \overline{X}_2\right)\frac{2 S_{cumulat} a}{\left(a^t S_{cumulat} a\right)^{3/2}} = 0 \qquad (22)$$

După o simplificare suplimentară obținem:

.

$$\overline{X}_1 - \overline{X}_2 = \left[\frac{a^t\left(\overline{X}_1 - \overline{X}_2\right)}{a^t S_{cumulat} a}\right] S_{cumulat}\, a \qquad (23)$$

Înmulțită cu inversul matricei S_{pooled} pe cele două părți obținem:

$$\overline{X}_1 - \overline{X}_2 = \left[\frac{a^t\left(\overline{X}_1 - \overline{X}_2\right)}{a^t S_{cumulat} a}\right] a\,, \qquad (24)$$

Deoarece $\dfrac{a^t(\overline{X}_1 - \overline{X}_2)}{a^t S_{cumulat} a}$ este un număr real obținem :

$$\hat{a} = c\, S_{cumulat}^{-1}\left(\overline{X}_1 - \overline{X}_2\right) \qquad (25)$$

unde c este o constantă.

- ***Metoda de clasificare***

Să presupunem că avem o observație notată cu X_0. Apoi, pe baza funcției discriminant $l(X) = \hat{a}^t X$ putem aloca această observație unei clase.

Rezultat important:

Alocarea lui X_0 a populației 1 în cazul în care

$$\hat{Y}_0 = \hat{a}^t X_0 = \left(\overline{X}_1 - \overline{X}_2\right)^t S_{pooled}^{-1} X_0 \geq \frac{1}{2}\hat{a}^t(\overline{X}_1 + \overline{X}_2) = \frac{1}{2}(\overline{Y}_1 + \overline{Y}_2)$$

$$= \frac{1}{2}\left(\overline{X}_1 - \overline{X}_2\right)^t S_{pooled}^{-1}\left(\overline{X}_1 + \overline{X}_2\right) \qquad (26)$$

Altfel, dacă $\hat{Y}_0 = (\overline{X}_1 - \overline{X}_2)^t S_{pooled}^{-1} X_0 < \frac{1}{2}\left(\overline{X}_1 - \overline{X}_2\right)^t S_{pooled}^{-1}\left(\overline{X}_1 + \overline{X}_2\right)$, atunci alocăm X_0 populației 2. În figura 2. este prezentată clasificarea populației 1 și 2.

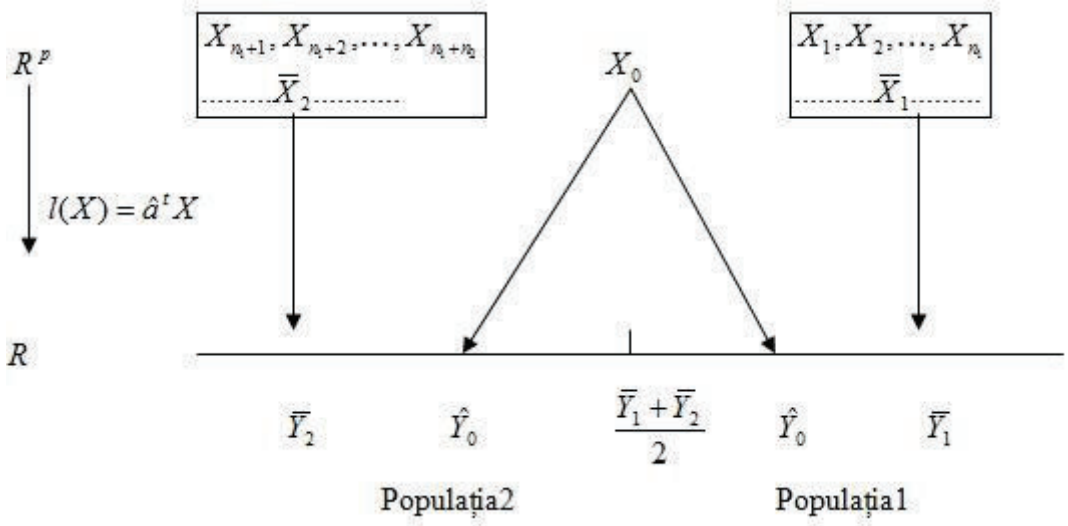


Figura 2. Reprezentarea clasificării celor 2 populații

Dacă $\hat{Y}_0$ se află pe partea dreaptă de $\frac{\overline{Y}_1 + \overline{Y}_2}{2}$, atunci se alocă X_0 populației 1 și invers.

Bibliografie:

1. Bartlett, M.S.,1954.*A note on themultiplyingfactors for various χ^2 approximations*. J. Roy. Statist. Soc. B 16, pp.296 -298.
2. Rao, C.R.,1973.*Linear StatisticalInferenceandIts Applications*.2nd Ed. Wiley, New York.