Lineární diskriminační analýza

Motivace:

Diskriminační analýza patří k vícerozměrným statistickým metodám a zabývá se klasifikací objektů do $r \geq 2$ skupin na základě znalosti vektorů pozorování těchto objektů. Zakladatelem DA je R. A. Fisher

Uveďme příklad z technické praxe: u výrobku daného typu potřebujeme rozhodnout, zda snese určitou zátěž. Na výrobku můžeme změřit hodnoty p kvantitativních znaků, např. hmotnost, odchylky rozměrů od normy, chemické složení apod., které tvoří vektor pozorování $\mathbf{x} = (x_1, ..., x_p)'$. Jestliže výrobek vystavíme zátěži, může to znamenat jeho poškození nebo dokonce zničení. Proto vystavíme zátěži jen omezené množství výrobků, řekněme n výrobků, které tvoří tzv. informativní výběr. Pokud výrobek zátěž vydrží, zařadíme ho do 1. skupiny (nechť takových výrobků je n_1) , jinak do 2. skupiny (těchto výrobků je n_2). Na základě znalosti chování informativního výběru pak rozložíme prostor $\mathbf{R}_p$ na množiny B_1, B_2. Máme-li k dispozici nějaký další výrobek téhož druhu s vektorem pozorování $\mathbf{x}$, zařadíme ho do 1. skupiny, když $\mathbf{x} \subset B_1$ a do 2. skupiny, když $\mathbf{x} \subset B_2$. Výrobek tedy nemusíme vystavovat zátěži a riskovat jeho poškození nebo dokonce zničení.

Úkol diskriminační analýzy spočívá v nalezení takového rozkladu prostoru $\mathbf{R}_p$ na množiny $B_1, \ldots, B_r$, který umožní optimální rozhodnutí o příslušnosti objektu ke skupině.

Náhodný výběr z vícerozměrného rozložení

Nechť je dáno n objektů a na každém z těchto objektů měříme p znaků. Znamená to, že i-tý objekt je charakterizován p-rozměrným vektorem pozorování $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})'$, který považujeme za realizaci náhodného vektoru

$\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})'$, i = 1, …, n. Všechny vektory pozorování uspořádáme do datové matice typu n x p: $\begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}$.

Předpokládáme, že náhodný vektor $\mathbf{X}_i$ má
vektor středních hodnot

$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_p \end{pmatrix}$

a varianční matici

$\boldsymbol{\Sigma} = (\sigma_{\cdot\cdot})$.

Lze dokázat, že nestranným odhadem vektoru $\boldsymbol{\mu}$ je vektor výběrových průměrů

$\mathbf{M} = \begin{pmatrix} M_1 \\ \vdots \\ M_p \end{pmatrix}$, kde $M_j = \frac{1}{n}\sum_{i=1}^{n} X_{ij}$ je výběrový průměr j-tého znaku, j = 1, …, p

a nestranným odhadem matice $\boldsymbol{\Sigma}$ je výběrová varianční matice

$\mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n} (\mathbf{X}_i - \mathbf{M})(\mathbf{X}_i - \mathbf{M})'$ řádu p.

Testy hypotéz o variančních maticích a vektorech středních hodnot

Nechť jsou dány dva p-rozměrné náhodné výběry o rozsazích n_1 a n_2 z p-rozměrných normálních rozložení $N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ a $N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$. Označme $\mathbf{M}_1$, $\mathbf{M}_2$ vektory výběrových průměrů, $\mathbf{S}_1$, $\mathbf{S}_2$ výběrové varianční matice a $\mathbf{S}$ vážený průměr výběrových variančních matic, tj. $\mathbf{S} = \frac{n_1 - 1}{n_1 + n_2 - 2}\mathbf{S}_1 + \frac{n_2 - 1}{n_1 + n_2 - 2}\mathbf{S}_2$.

a) **Test shody variančních matic (Boxův test)**

Testujeme hypotézu H_0: $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$ proti alternativní hypotéze H_1: $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$ na hladině významnosti α. Test je založen na Boxově statistice:

$M = (n_1 + n_2 - 2) \ln (\det \mathbf{S}) - (n_1 - 1) \ln (\det \mathbf{S}_1) - (n_2 - 1) \ln (\det \mathbf{S}_2)$.

Označme konstantu $C_p = 1 - \frac{2p^2 + 3p - 1}{6(p+1)}\left(\frac{1}{n_1 - 1} + \frac{1}{n_2 - 1} - \frac{1}{n_1 + n_2 - 2}\right)$. (Tato konstanta zlepšuje aproximaci.)

Platí-li H_0, pak testová statistika MC_p má asymptoticky rozložení $\chi^2\left(\frac{p(p+1)}{2}\right)$.

Kritický obor: $W = \left\langle \chi^2_{1-\alpha}\left(\frac{p(p+1)}{2}\right), \infty \right)$.

Pokud $MC_p \in W$, H_0 zamítáme na asymptotické hladině významnosti α.

b) **Test shody vektorů středních hodnot**

Nezamítneme-li na zvolené hladině významnosti hypotézu o shodě variančních matic, můžeme přistoupit k testování hypotézy H_0: $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ proti alternativní hypotéze H_1: $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$ na hladině významnosti α.

Označme $T^2 = \frac{n_1 n_2}{n_1 +}(\mathbf{M}_1 - \mathbf{M}_2)' \mathbf{S}^{-1}(\mathbf{M}_1 - \mathbf{M}_2)$.

Nulovou hypotézu zamítneme na hladině významnosti α, když testová statistika

$\frac{n_1 + \quad -}{p(n_1 + \quad -} \quad ^2 \geq F_{1-\alpha}(p, n_1+n_2-p-1)$.

Poznámka:

Zamítneme-li hypotézu o shodě vektorů středních hodnot, je vhodné provést testy dílčích hypotéz

H_{0j}: $\mu_{j1} = \mu_{j2}$ proti H_{1j}: $\mu_{j1} \neq \mu_{j2}$, j =1, ..., p. K tomu slouží dvouvýběrový t-test.

Hypotézu H_{0j} zamítneme na hladině významnosti α ve prospěch oboustranné alternativy, když

$|T_i| = \frac{|M_{j1} - \quad _{j2}|}{\sqrt{s_{jj}\left(\frac{1}{1} + \right)}} \geq - \left(n_1 + \quad -\right)$, kde s_{jj} je j-tý diagonální prvek matice **S**.

Odvození bayesovského rozhodovacího pravidla pro dvě skupiny objektů

Nechť v 1. skupině je n_1 objektů, ve 2. skupině n_2 objektů, přičemž každý objekt je charakterizován p-rozměrným náhodným vektorem $\mathbf{X} = (X_1, \ldots, X_p)'$.

Předpokládáme, že v i-té skupině má náhodný vektor $\mathbf{X}$ hustotu $\varphi_i(\mathbf{x})$, i = 1, 2.

Nechť H_i je jev „objekt patří do i-té skupiny“.

Apriorní pravděpodobnost $P(H_i)$ příslušnosti objektu k i-té skupině označíme π_i, i = 1, 2.

Známe-li u nějakého objektu vektor pozorování $\mathbf{x}$, můžeme podle Bayesova vzorce vypočítat aposteriorní pravděpodobnost příslušnosti objektu ke skupině:

$P(H_i/\mathbf{X}=\mathbf{x}) = \frac{\pi_i \varphi_i(\mathbf{x})}{\pi_1 \varphi_1(\mathbf{x}) + \pi_2 \varphi_2(\mathbf{x})}$, i = 1, 2.

Nabízí se jednoduché rozhodovací pravidlo: zařadit nový objekt do té skupiny, u níž je aposteriorní pravděpodobnost větší.

Tedy objekt s vektorem pozorování $\mathbf{x}$ zařadíme do 1. skupiny, když $\pi_1\varphi_1(\mathbf{x}) > \pi_2\varphi_2(\mathbf{x})$, jinak ho zařadíme do 2. skupiny.

Součin $\pi_i\varphi_i(\mathbf{x})$ se nazývá diskriminační skór pro i-tou skupinu.

Lze ukázat, že bayesovské rozhodovací pravidlo je optimální v tom smyslu, že minimalizuje celkovou pravděpodobnost mylné klasifikace.

Konstrukce Fisherovy lineární diskriminační funkce pro dvě skupiny objektů

Předpokládejme nyní, že hustota pravděpodobnosti v i-té skupině je normální a má parametry $\boldsymbol{\mu}_i$, $\boldsymbol{\Sigma}_i$, tj.

$$\varphi_i(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^p \det \boldsymbol{\Sigma}_i}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)' \boldsymbol{\Sigma}_i^{-1} (\mathbf{x}-\boldsymbol{\mu}_i)\right), \quad i = 1, 2.$$

Jestliže zlogaritmujeme diskriminační skór $\pi_i\varphi_i(\mathbf{x})$ a vynecháme člen $-\frac{p}{2}\ln(2\pi)$, který je společný pro obě skupiny,

dostaneme tzv. kvadratický diskriminační skór pro i-tou skupinu ve tvaru $-\frac{1}{2}\ln\det\boldsymbol{\Sigma}_i - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)'\boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) + \ln\pi_i$, $i = 1, 2$.

Jsou-li varianční matice v obou skupinách stejné (společnou varianční matici označíme $\boldsymbol{\Sigma}$), obsahují oba kvadratické diskriminační skóry týž člen $-\frac{1}{2}\ln\det\boldsymbol{\Sigma} - \frac{1}{2}\mathbf{x}'\boldsymbol{\Sigma}^{-1}\mathbf{x}$. Po jeho vynechání obdržíme lineární diskriminační skór pro i-tou skupinu

– tzv. Andersonovu diskriminační statistiku - ve tvaru $\lambda_i(\mathbf{x}) = \boldsymbol{\mu}_i'\boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_i'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i + \ln\pi_i$, $i = 1, 2$.

Objekt s vektorem pozorování $\mathbf{x}$ tedy zařadíme do 1. skupiny, když $\lambda_1(\mathbf{x}) > \lambda_2(\mathbf{x})$, jinak ho zařadíme do 2. skupiny.
Vzhledem k tomu, že máme jen dvě skupiny objektů, lze rozhodnutí o zařazení objektu do skupiny učinit na základě rozdílu

$\lambda(\mathbf{x}) = \lambda_1(\mathbf{x}) - \lambda_2(\mathbf{x}) = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_1'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_2'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2 + \ln\pi_1 - \ln\pi_2$.

Funkce $\lambda(\mathbf{x})$ se nazývá Fisherova lineární diskriminační funkce. Označíme-li

$\boldsymbol{\beta}' = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}$, $\gamma = -\frac{1}{2}\boldsymbol{\beta}'(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) + \ln\pi_1 - \ln\pi_2$,

můžeme Fisherovu lineární diskriminační funkci psát ve tvaru
$\lambda(\mathbf{x}) = \boldsymbol{\beta}'\mathbf{x} + \gamma$.
Znamená to, že jsme našli takovou lineární kombinaci vektoru pozorování $\mathbf{x}$, která nám umožní minimalizovat celkovou pravděpodobnost mylného zařazení objektu do skupiny. Objekt s vektorem pozorování $\mathbf{x}$ tedy zařadíme do 1. skupiny, když $\lambda(\mathbf{x}) > 0$, jinak ho zařadíme do 2. skupiny.

Posouzení účinnosti diskriminace resubstituční metodou

Resubstituční metoda spočívá v uplatnění zkonstruovaného rozhodovacího pravidla na informativní výběr. Uvažujeme postupně všechny objekty z informativního výběru a jejich zařazení podle rozhodovacího pravidla porovnáme se skutečnou příslušností ke skupině. Stanovíme podíl správně a mylně zařazených objektů.

skutečnost	zařazení		součet
	1. skupina	2. skupina	
1. skupina	n_{11}	n_{12}	$n_{1.} = n_1$
2. skupina	n_{21}	n_{22}	$n_{2.} = n_2$
součet	$n_{.1}$	$n_{.2}$	n

Podíl správně zařazených objektů:

$$\frac{n_{11} + \quad}{n}$$

Podíl mylně zařazených objektů:

$$\frac{n_{12} + \quad}{n}$$

Modifikace pro případ neznámých parametrů

Při praktickém použití diskriminační analýzy většinou neznáme parametry $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, $\boldsymbol{\Sigma}$ ani apriorní pravděpodobnosti π_1, π_2.

V takovém případě používáme odhady:

$\boldsymbol{\mu}_i \rightarrow \mathbf{M}_i$, i = 1, 2

$\boldsymbol{\Sigma} \rightarrow \mathbf{S} = \frac{1}{n_1 + n_2 - 2}(n_1 \mathbf{S}_1 + n_2 \mathbf{S}_2)$

$\pi_i \rightarrow \frac{n_i}{n}$, i = 1, 2.

Odhad Fisherovy lineární diskriminační funkce $\lambda(\mathbf{x}) = \boldsymbol{\beta}'\mathbf{x} + \gamma$:

$L(\mathbf{x}) = \mathbf{b}'\mathbf{x} + g$, kde

$\mathbf{b}' = (\mathbf{M}_1 - \mathbf{M}_2)'\mathbf{S}^{-1}$, $g = -\frac{1}{2}\mathbf{b}'(\mathbf{M}_1 + \mathbf{M}_2) + \ln p_1 - \ln p_2$.

Postup při lineární diskriminační analýze

1. Vzhledem k povaze úlohy určíme veličiny $X_1, ..., X_p$ a pořídíme $n_1 + n_2$ p-rozměrných pozorování tak, aby n_1 objektů pocházelo z 1. skupiny a n_2 objektů z 2. skupiny.
2. Na zvolené hladině významnosti α testujeme hypotézy o normalitě rozložení v obou skupinách.
3. Vypočteme odhady $\mathbf{M}_1$, $\mathbf{M}_2$, $\mathbf{S}_1$, $\mathbf{S}_2$, $\mathbf{S}$, p_1, p_2.
4. Na zvolené hladině významnosti α testujeme hypotézy o shodě variančních matic a vektorů středních hodnot v obou skupinách.
5. Vypočteme odhad $L(\mathbf{x})$ Fisherovy lineární diskriminační funkce. Objekt s vektorem pozorování $\mathbf{x}$ přiřadíme k 1. skupině, když $L(\mathbf{x}) > 0$, jinak ho přiřadíme ke 2. skupině.
6. Účinnost diskriminace posoudíme metodou resubstituce.

Požadavky na vstupní soubor dat:

a) charakteristické znaky jednotlivých prvků musí být kvantitativní,

b) žádný ze znaků nesmí být lineární kombinací ostatních znaků (lineární nezávislost),

V diskriminační analýze musí pro počty skupin, počty znaků, počty objektů ve skupinách a celkové počty objektů platit:

a) skupiny objektů musí být minimálně dvě,

b) každá skupina musí mít alespoň dva objekty,

c) počet znaků použitých v analýze musí být menší než počet objektů zmenšený o počet skupin,

d) žádný znak by neměl být konstantní v jakékoliv skupině.

Příklad

V souboru 50 rodin byly zjišťovány tyto údaje:

- zda v posledních dvou letech rodina navštívila jistou rekreační oblast (veličina ID, nabývá hodnoty 0 pro odpověď „ne“, hodnoty 1 pro odpověď „ano“)
- roční příjem v tisících dolarů (veličina X_1)
- postoj k cestování (veličina X_2, devítibodová škála, 1 = naprosto odmítavý, 9 = veskrze kladný)
- význam přičítaný rodinné dovolené (veličina X_3, devítibodová škála, 1 = nejnižší, 9 = nejvyšší)
- počet členů rodiny (veličina X_4)
- věk nejstaršího člena rodiny (veličina X_5).

Pro uvedená data sestrojte Fisherovu lineární diskriminační funkci, která pomocí veličin X_1, …, X_5 umožní rozlišit rodiny navštěvující uvedenou rekreační oblast od rodin, které do této oblasti nejezdí.

Datový soubor:

číslo	ID	X_1	X_2	X_3	X_4	X_5	číslo	ID	X_1	X_2	X_3	X_4	X_5
1.	0	32,1	5	4	6	58,0	26.	0	48,2	3	5	4	43,0
2.	0	40,0	4	4	3	42,0	27.	0	54,5	7	3	3	37,0
3.	0	36,2	4	3	2	55,0	28.	0	38,2	2	5	3	49,0
4.	0	43,2	2	5	2	57,0	29.	0	41,7	4	2	3	40,0
5.	0	50,4	5	2	4	37,0	30.	1	50,2	5	8	3	43,0
6.	0	45,2	4	4	4	42,0	31.	1	70,3	6	7	4	61,0
7.	0	44,1	6	6	3	42,0	32.	1	62,9	7	5	6	52,0
8.	0	38,3	6	6	2	45,0	33.	1	48,5	7	5	5	36,0
9.	0	55,0	1	5	4	57,0	34.	1	52,7	6	6	4	55,0
10.	0	56,1	3	5	5	51,0	35.	1	75,0	8	7	5	68,0
11.	0	48,2	4	3	6	47,0	36.	1	46,2	5	3	3	62,0
12.	0	35,0	6	4	5	64,0	37.	1	57,0	2	4	6	51,0
13.	0	37,3	2	7	3	54,0	38.	1	64,1	4	5	4	57,0
14.	0	41,8	5	1	5	56,0	39.	1	68,1	4	6	5	45,0
15.	0	57,0	8	3	4	36,0	40.	1	73,4	6	7	5	44,0
16.	0	33,4	6	8	4	50,0	41.	1	71,6	5	8	4	64,0
17.	0	41,5	5	6	3	38,0	42.	1	56,2	1	8	6	54,0
18.	0	39,8	4	5	4	42,0	43.	1	49,3	4	2	3	56,0
19.	0	37,5	3	2	3	48,0	44.	1	62,0	5	6	2	58,0
20.	0	41,3	3	3	2	42,0	45.	1	50,8	4	7	3	45,0
21.	0	35,0	4	3	4	54,0	46.	1	63,6	7	4	7	55,0
22.	0	49,6	5	5	5	39,0	47.	1	54,0	6	7	4	58,0
23.	0	45,5	4	4	4	41,0	48.	1	49,0	5	4	3	60,0
24.	0	39,4	6	5	3	44,0	49.	1	68,0	6	6	6	46,0
25.	0	37,0	2	6	5	51,0	50.	1	62,1	5	6	3	56,0

Řešení:

Testování normality náhodných veličin $X_1, \ldots, X_5$ v daných dvou skupinách rodin pomocí S - W testu:

Pro skupinu rodin, které danou rekreační oblast nenavštěvují: Statistiky – Základní statistiky/tabulky – Select cases – ID=0 – OK – Tabulky četností – Proměnné X1 až X5 – OK – Normalita – zaškrtneme S-W test – Testy normality

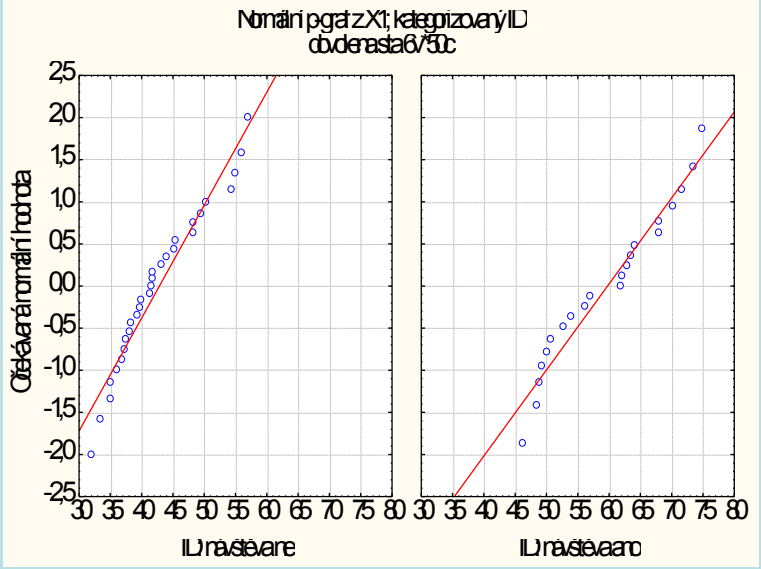
Proměnná	Testy normality (d Zhrnout podmínku N	W	p
X1: roční příjem v tisících dola	2	0,940	0,101
X2: postoj k cestování (škála 9	2	0,964	0,412
X3: význam rodinné dovolené	2	0,964	0,420
X4: počet členů rodiny	2	0,917	0,026
X5: věk nejstaršího člena	2	0,944	0,131

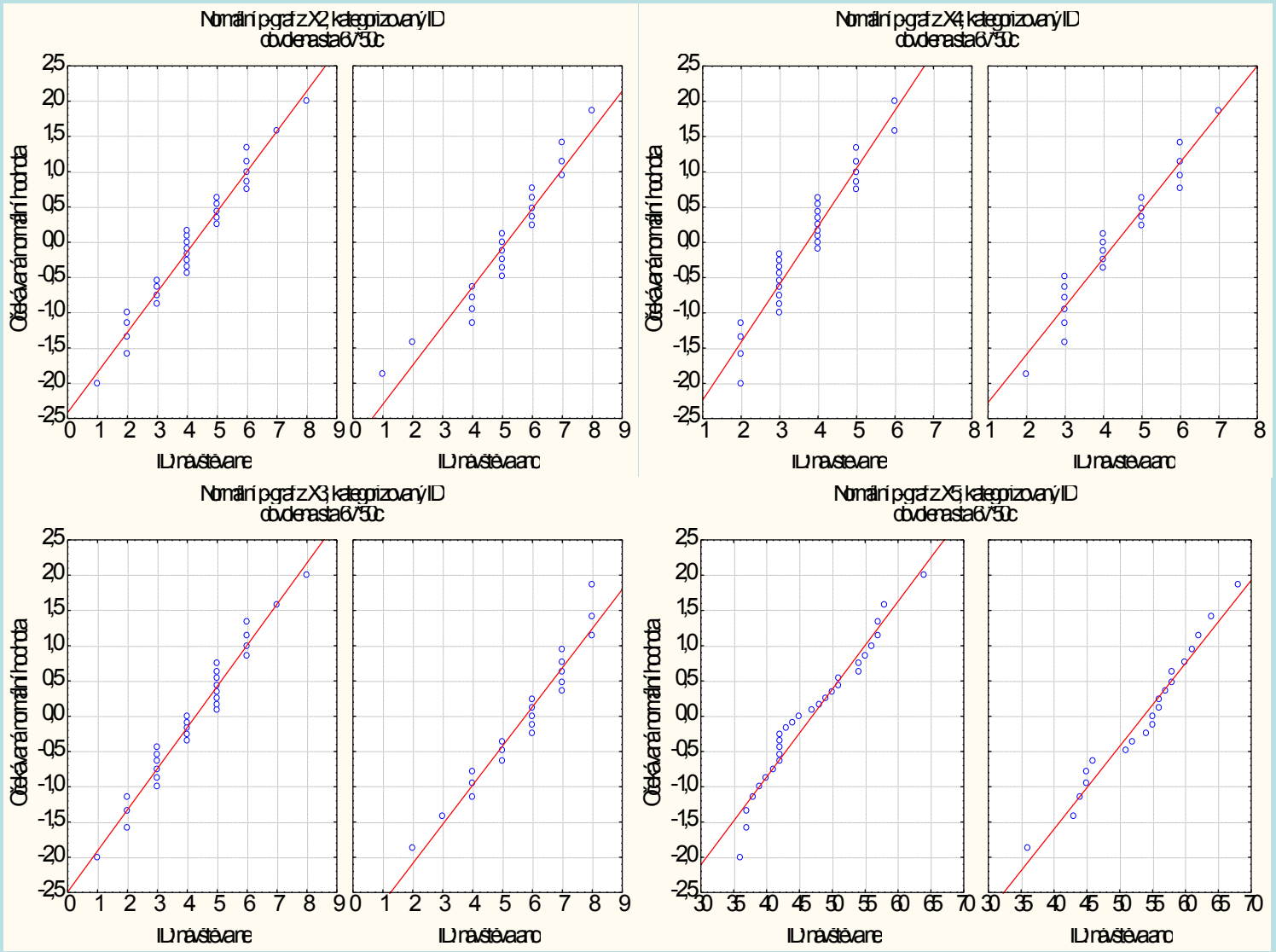
Pro skupinu rodin, které danou rekreační oblast navštěvují: Statistiky – Základní statistiky/tabulky – Select cases – ID=1 – OK – Tabulky četností – Proměnné X1 až X5 – OK – Normalita – zaškrtneme S-W test – Testy normality

Proměnná	Testy normality (d Zhrnout podmínku N	W	p
X1: roční příjem v tisících dola	2	0,935	0,180
X2: postoj k cestování (škála 9	2	0,930	0,139
X3: význam rodinné dovolené	2	0,934	0,171
X4: počet členů rodiny	2	0,928	0,126
X5: věk nejstaršího člena	2	0,967	0,679

Na hladině významnosti 0,05 zamítáme hypotézu o normalitě u veličiny X_4 ve skupině rodin, které danou rekreační oblast nenavštěvují.

N-P ploty:
Grafy – 2D Grafy – Normální pravděpodobnostní grafy – Proměnné X1 až X5 – OK – na záložce Kategorizovaný zaškrtneme Kategorie X Zapnuto – Změnit proměnnou – ID – OK – OK



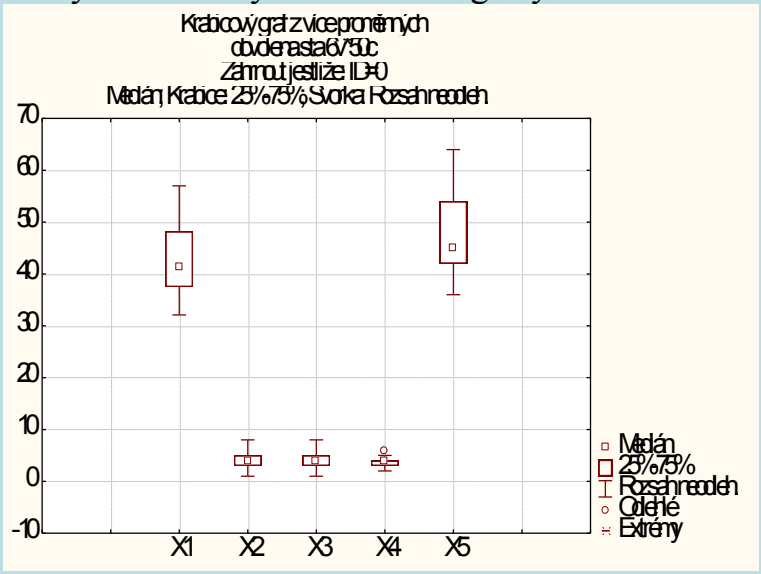

Normální p-graf z X2, kategorizovaný ID
obvderasta6/*30c
Očekávaná normální hodnota
ID návštěvare
ID návštěvaarc
Normální p-graf z X4, kategorizovaný ID
obvderasta6/*30c
Očekávaná normální hodnota
ID návštěvare
ID návštěvaarc
Normální p-graf z X3, kategorizovaný ID
obvderasta6/*30c
Očekávaná normální hodnota
ID návštěvare
ID návštěvaarc
Normální p-graf z X5, kategorizovaný ID
obvderasta6/*30c
Očekávaná normální hodnota
ID návštěvare
ID návštěvaarc

Odhad vektorů středních hodnot $\mathbf{M}_1$ a $\mathbf{M}_2$ lze získat více způsoby, uvedeme např. tento:
Statistiky – Základní statistiky/tabulky – Select cases – ID=0 - Popisné statistiky – Proměnné X1 až X5 – Grupovací proměnná ID=0 – OK – Detailní výsledky – zaškrtneme pouze N a průměr – Souhrn

Popisné statistiky (d
Zhrnout podmínku:

Proměn	N platn	Průmě
X1	2	42,84
X2	2	4,241
X3	2	4,275
X4	2	3,724
X5	2	46,93

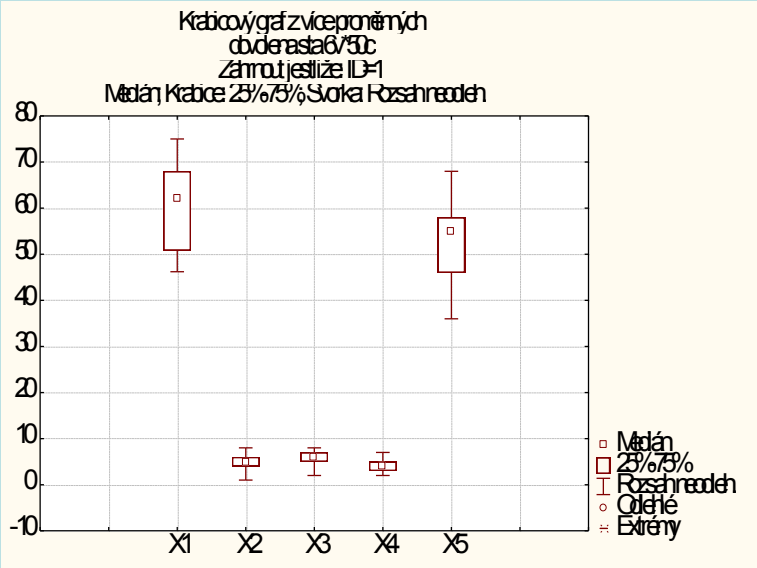
Krabicové grafy:
Grafy – 2D Grafy – Krabicové grafy – Vícenásobný – Závisle proměnné X1 až X5 – OK – OK



Nyní změníme podmínku ID = 1

Popisné statistiky (d
Zhrnout podmínku:

Proměn	N platn	Prům
X1	2	59,76
X2	2	5,142
X3	2	5,76
X4	2	4,333
X5	2	53,61



Odhad varianční matice $\mathbf{S}_1$:

Statistiky – Vícerozměrná regrese – Select cases ID=0 – OK – Proměnné - Závislá proměnná X5, Seznam nezávisle proměnných X1 až X4 – OK – OK - Residua/předpoklady/předpovědi – Popisné statistiky – Další statistiky - Kovariance

Kovariance (dovolena.sta) Zhrnout podmínku: ID=0					
Proměn	X1	X2	X3	X4	X5
X1	49,19	0,995	-2,24	1,094	-24,1(
X2	0,99	2,76(	-0,31(	0,140	-4,7(
X3	-2,24	-0,31(	2,63(	-0,171	1,12
X4	1,09	0,14(	-0,17	1,278	1,94
X5	-24,1(	-4,73(	1,12(	1,944	57,28

Odhad varianční matice $\mathbf{S}_2$: Změníme podmínku ID=1

Kovariance (dovolena.sta) Zhrnout podmínku: ID=1					
Proměn	X1	X2	X3	X4	X5
X1	83,59	4,300	6,39(	4,70(	16,25
X2	4,30(	2,728	0,03(	0,20(	1,05(
X3	6,39(	0,035	2,79(	0,03(	-1,04(
X4	4,70(	0,200	0,03(	1,83(	-2,46(
X5	16,25	1,057	-1,04(	-2,46(	63,84

Odhad společné varianční matice **S**:
Statistiky – Vícerozměrné průzkumné techniky – Diskriminační analýza – Proměnné – Grupovací ID, Seznam nezáv. proměnných X1-X5 – OK, zapneme Další možnosti (kroková analýza) – OK – Popisné statistiky – Zobrazit popisné statistiky – Vnitřní kovariance a korelace.

	X_1	X_2	X_3	X_4	X_5
X1	63,[illegible]	2,[illegible]	1,[illegible]	2,[illegible]	-7,[illegible]
X2	2,[illegible]	2,[illegible]	-0,[illegible]	0,[illegible]	-2,[illegible]
X3	1,[illegible]	-0,[illegible]	2,[illegible]	-0,[illegible]	0,[illegible]
X4	2,[illegible]	0,[illegible]	-0,[illegible]	1,[illegible]	0,[illegible]
X5	-7,[illegible]	-2,[illegible]	0,[illegible]	0,[illegible]	60,[illegible]

Boxův test shody variančních matic:

Statistika $M = (n_1 + n_2 - 2) \ln(\det \mathbf{S}) - (n_1 - 1) \ln(\det \mathbf{S}_1) - (n_2 - 1) \ln(\det \mathbf{S}_2) = 26{,}6179$

Konstanta zlepšující aproximaci $C_p = 1 - \frac{2p^2 + 3p - 1}{6(p+1)} \left(\frac{1}{n_1 - 1} + \frac{1}{n_2 - 1} - \frac{1}{n_1 + n_2 - 2} \right) = 0{,}8847$

Testová statistika $MC_p = 23{,}5468$

Kritický obor: $W = \left\langle \chi^2_{1-\alpha}\left(\frac{p(p+1)}{2}\right), \infty \right\rangle = \langle 10{,}24958, \infty \rangle$.

Protože testová statistika neleží v kritickém oboru, nezamítáme na asymptotické hladině významnosti 0,05 hypotézu o shodě variančních matic $\mathbf{\Sigma}_1$, $\mathbf{\Sigma}_2$.

Provedení testu v systému STATISTICA:

Statistiky – Pokročilé lineární/nelineární modely – Obecné lineární modely – Typ analýzy: Jednofaktorová ANOVA - Metoda specifikace: Rychlé nastavení – OK – Proměnné – Seznam závislých proměnných: X1 – X5, Kategor. nezávislá proměnná (faktor): ID – OK – OK – Více výsledků – na záložce Předpoklady vybereme Boxův M-test.

	Boxův M test (dovolen Efekt: ID (Vypočteno pro všechr			
	Boxov(	Chi-k	s\	p
Boxovc	26,61	23,54	1:	0,073

Protože p-hodnota je větší než hladina významnosti 0,05, hypotézu o shodě variančních matic nezamítáme na asymptotické hladině významnosti 0,05.

Test shody vektorů středních hodnot:

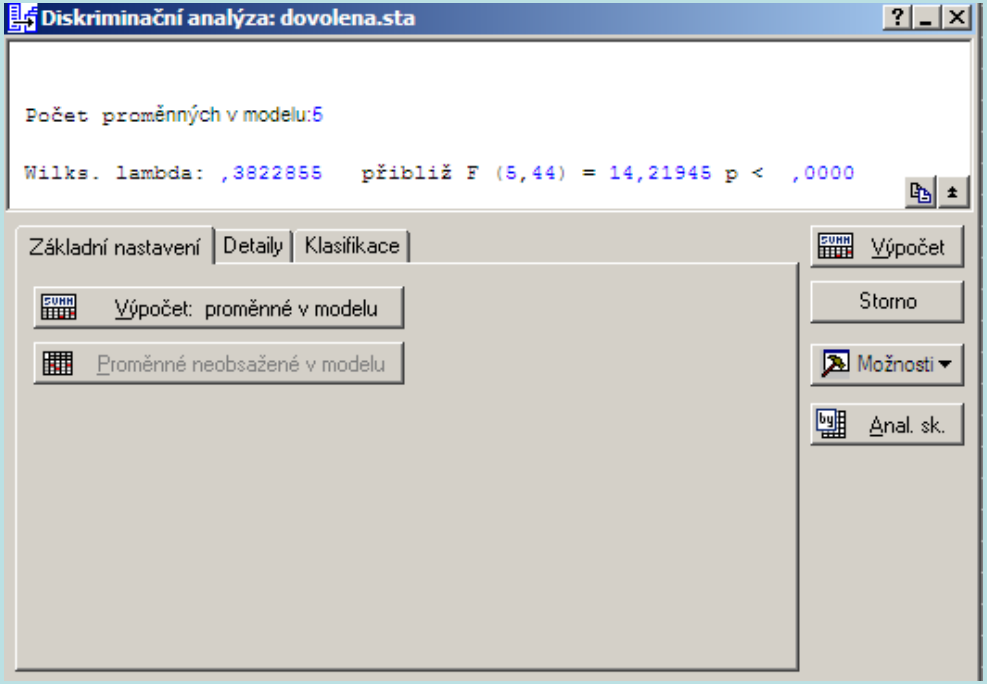
Testová statistika $\frac{n_1 + n_2 - p - 1}{p(n_1 + n_2 - 2)} \cdot \frac{n_1 n_2}{n_1 + n_2} (\mathbf{M}_1 - \mathbf{M}_2)' \mathbf{S}^{-1} (\mathbf{M}_1 - \mathbf{M}_2) = 14{,}2194$

Kvantil $F_{1-\alpha}(p, n_1+n_2-p-1) = F_{0,95}(5,44) = 2{,}427$

Protože testová statistika se realizuje v kritickém oboru, zamítáme na hladině významnosti 0,05 hypotézu o shodě vektorů středních hodnot $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$.

Výpočet testové statistiky v systému STATISTICA:

Statistiky – Vícerozměrné průzkumné techniky – Diskriminační analýza - Proměnné – Grupovací proměnná ID, Seznam nezávislých proměnných X1 až X5 — OK – OK



Upozornění: Test shody vektorů středních hodnot lze v systému STATISTICA provést i jinak:
Statistiky – Základní statistiky/tabulky – t-test, nezávislé, dle skupin – OK – Proměnné – Závisle proměnné X1 až X5, Grupovací proměnná ID – OK – na záložce Možnosti zaškrtneme Vícerozměrný test. V záhlaví výstupní tabulky se zobrazí realizace testové statistiky a příslušná p-hodnota.

t-testy; grupováno: ID (dovolena.sta)
Skup. 1: návštěva ne; Skup. 2: návštěva ano
Hotellingovo 77,5606 F(5,44)=14,219 p<,00000

Proměn	Průměr návštěv	Průměr návštěva	t	sv	p	Poč.pla návštěv	Poč.pla návštěva	Sm.odc návštěv	Sm.odc návštěva
X1	42,84	59,76	-7,40	4	0,000	2	2	7,013	9,142
X2	4,24	5,142	-1,89	4	0,063	2	2	1,661	1,651
X3	4,27	5,76	-3,15	4	0,002	2	2	1,623	1,670
X4	3,724	4,333	-1,73	4	0,089	2	2	1,130	1,354
X5	46,93	53,61	-3,01	4	0,004	2	2	7,568	7,990

Vidíme, že na hladině významnosti 0,05 jsou odlišné střední hodnoty proměnných X_1, X_3, X_5. U proměnných X_2 a X_4 se odlišnost neprokázala, z dalšího zpracování je však vyřazovat nebudeme.

Stanovení odhadů apriorních pravděpodobností:

$p_1 = \frac{n_1}{n} = \frac{29}{50} = 0{,}58$ $p_2 = \frac{n_2}{n} = \frac{21}{50} = 0{,}42$

Stanovení odhadu Fisherovy lineární diskriminační funkce:

$\mathbf{b}' = (\mathbf{M}_1 - \mathbf{M}_2)'\mathbf{S}^{-1} = (-0{,}2865 \;\; -0{,}2556 \;\; -0{,}4169 \;\; 0{,}0736 \;\; -0{,}1527)$

$g = -\frac{1}{2}\mathbf{b}'(\mathbf{M}_1 + \mathbf{M}_2) + \ln p_1 - \ln p_2 = 24{,}7666$

$L(\mathbf{x}) = \mathbf{b}'\mathbf{x} + g = -0{,}2685X_1 - 0{,}2556X_2 - 0{,}4169X_3 + 0{,}0736X_4 - 0{,}1527X_5 + 24{,}7666$

Statistiky – Vícerozměrné průzkumné techniky – Diskriminační analýza - Proměnné – Grupovací proměnná ID, Seznam nezávislých proměnných X1 až X5 — OK – OK – na záložce Klasifikace zvolíme Klasifikační funkce. Dostaneme tabulku tvaru:

	Klasifikační funkce; grupov	
Proměnná	návštěva p=,580	návštěva p=,420
X1	0,63	0,90
X2	1,78	2,03
X3	1,33	1,75
X4	1,18	1,11
X5	0,92	1,07
Konstanta	-44,6	-69,4

Abychom získali odhad Fisherovy lineární diskriminační funkce, přidáme do této tabulky novou proměnnou a do jejího Dlouhého jména napíšeme =v1-v2

	Klasifikační funkce; grupov		
Proměnná	návštěva p=,580	návštěva p=,420	NPro =v1-v
X1	0,63	0,90	-0,26
X2	1,78	2,03	-0,25
X3	1,33	1,75	-0,41
X4	1,18	1,11	0,073
X5	0,92	1,07	-0,15
Konstanta	-44,6	-69,4	24,76

Posouzení účinnosti diskriminace resubstituční metodou:

skutečnost	zařazení		součet
	návštěva ne	návštěva ano	
návštěva ne	27	2	29
návštěva ano	5	16	21
součet	32	18	50

Podíl správně zařazených objektů:

$$\frac{n_{11}+n_{22}}{n} = \frac{27+16}{50} = 0,86$$

Podíl mylně zařazených objektů:

$$\frac{n_{12}+n_{21}}{n} = \frac{2+5}{50} = 0,14$$

Na záložce Klasifikace zvolíme Klasifikační matice.

Skup.	Klasifikační matice (dovol Řádky: pozorované klasifi Sloupce: předpovězené kl % správn	návštěva p=,580	návštěva p=,420
návštěva	93,10	2	2
návštěva	76,19	5	1
Celkem	86,00	3	1

Pro určení chybně zařazených případů zvolíme na záložce Klasifikace možnost Klasifikace případů. Zjistíme, že v 1. skupině došlo k mylnému zařazení u rodin č. 9 a 10, ve 2. skupině u rodin číslo 30, 33, 36, 43, 45.

Klasifikace nového případu

Předpokládejme nyní, že jsme prozkoumali další rodinu, která má roční příjem 51,8 tisíc dolarů, k cestování zaujímá postoj ohodnocený 6 body, rodinné dovolené přičítá význam ohodnocený 7 body, má 4 členy a nejstaršímu členovi je 51 let. Na základě těchto údajů se pokusíme pomocí Fisherovy lineární diskriminační funkce zařadit tuto rodinu do skupiny rodin, které buď navštěvují nebo nenavštěvují danou rekreační oblast.

$L(\mathbf{x}) = -0,2685X_1 - 0,2556X_2 - 0,4169X_3 + 0,0736X_4 - 0,1527X_5 + 24,7666 =$

$= -0,2685.51,8 - 0,2556.6 - 0,4169.7 + 0,0736.4 - 0,1527.51 + 24,7666 = -1,0836.$

Protože $L(\mathbf{x}) < 0$, zařadíme tuto rodinu do skupiny rodin, které navštěvují danou rekreační oblast.

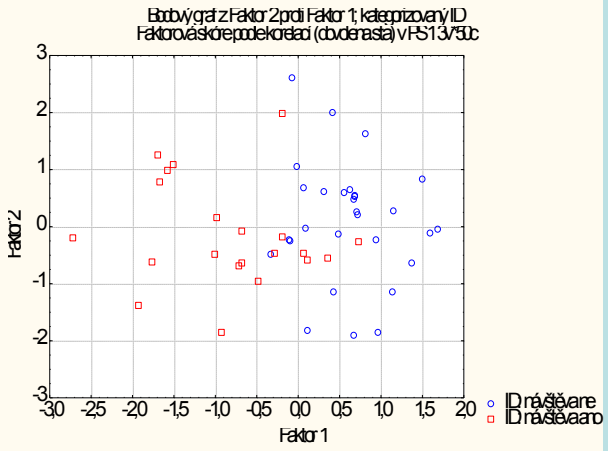
Porovnání s náhodnou klasifikací

Kdybychom zařazovali rodiny do skupin náhodně, pouze s ohledem na apriorní pravděpodobnosti π_1, π_2, tak bychom s pravděpodobností π_1 našli rodinu patřící do 1. skupiny, avšak s pravděpodobností π_2 bychom ji mylně zařadili do 2. skupiny. Naopak s pravděpodobností π_2 najdeme rodinu patřící do 2. skupiny, kterou s pravděpodobností π_1 mylně zařadíme do 1. skupiny. Celková pravděpodobnost mylné klasifikace je tedy: $\pi_1\pi_2 + \pi_2\pi_1 = 2\pi_1(1- \pi_1)$. Nahradíme-li apriorní pravděpodobnosti π_1, π_2 jejich odhady p_1, p_2 , dostaneme odhad celkové pravděpodobnosti mylné klasifikace $2p_1(1- p_1) = 2 \cdot \frac{29}{50} \cdot \frac{21}{50} = 0{,}4872$.

Použitím diskriminační analýzy jsme tedy dosáhli výrazného zlepšení, pravděpodobnost mylné klasifikace klesla na 0,14.

Grafické znázornění případů na ploše prvních dvou hlavních komponent

Jako aktivní vstup použijeme Faktorová skóre podle korelací z analýzy hlavních komponent. Grafy – Kategorizované grafy – Bodové grafy – Rozložení Přes sebe – Proměnné X: Faktor 1, Y: Faktor 2, X_Kategorie: ID - OK



Literatura

[1] J. Anděl: Matematická statistika, SNTL/Alfa, Praha 1978

[2] J. Anděl: Statistické metody, Matfyzpress, Praha 1993

[3] P. Hebák, J. Hustopecký, E. Jarošová, I. Pecáková: Vícerozměrné statistické metody (1), Informatorium, Praha 2004