Osnova přednášky Lineární diskriminační analýza

1. 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

Řeší problém, jak získat jednu či více rovnic, které umožní klasifikovat objekty do skupin. Tyto rovnice se nazývají klasifikační neboli diskriminační funkce a kombinují jednotlivé proměnné a jejich váhy tak, aby bylo možné určit skupinu, do které klasifikovaný objekt s největší pravděpodobností patří.

2. Možnosti použití diskriminační analýzy

Technické obory:
Při kontrole jakosti či spolehlivosti lze ve výběrovém souboru výrobků změřit nějaké kvantitativní proměnné (např. rozměry, hmotnost, chemické složení apod.), pak výrobky podrobit zátěži a sledovat, zda tuto zátěž vydrží nebo ne. K predikci chování dalších výrobků při zátěži je skutečné zátěži nemusíme vystavovat, stačí, když provedeme potřebná měření kvantitativních proměnných.

Lékařství
Máme soubor pacientů, u nichž jsou diagnostikovány určité choroby. Pro každého pacienta máme k dispozici výsledky různých laboratorních testů. Pokud existuje souvislost mezi výsledky testů a diagnózou, může se lékař u nových pacientů rozhodovat pro určitou diagnózu (a tedy i způsob léčení) na základě výsledků testů.

Bankovnictví
Banka sleduje ve výběrovém souboru klientů, jak splácejí poskytnutý úvěr a kromě toho řadu dalších ukazatelů (věk, rodinný stav, výši příjmu, …). Následně na tomto základě může vyhodnocovat potenciální žadatele o úvěr jako více či méně důvěryhodné.

Archeologie
Při vykopávkách byly nalézány hroby s kostrami pravěkých lidí. Na základě nějakých charakteristických vlastností (délka určité kosti, úhly kostí na lebce,…) bylo možné další nalezené kostry zařadit k určitému historickému období, kultuře a rase.

3. Opakování

3.1. 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: $\begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}$.

Náhodný vektor $\mathbf{X}_i$ má vektor středních hodnot $\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_p \end{pmatrix}$ a varianční matici $\boldsymbol{\Sigma} = (\sigma_{ij})_{i=1,j=1}^{p,p}$.

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

Nadále předpokládáme, že máme $r \geq 2$ p-rozměrných náhodných výběrů o rozsazích $n_1, \ldots, n_r$, přičemž h-tý výběr $\mathbf{X}_{h1}, \ldots, \mathbf{X}_{hn_h}$ pochází z p-rozměrného normálního rozložení $N_p(\boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h)$, $h = 1, \ldots, r$ a jednotlivé náhodné výběry jsou stochasticky nezávislé.

3.2. Boxův test shody variančních matic

Na hladině významnosti α testujeme hypotézu $H_0 : \mathbf{\Sigma}_1 = \ldots = \mathbf{\Sigma}_r$ proti alternativní hypotéze H_1 : aspoň jedna dvojice variančních matic se liší. Testová statistika má tvar:

$$T_0 = \frac{1}{C_p}\left[(n-r)\ln|\mathbf{S}| - \sum_{h=1}^{r}(n_h - 1)\ln|\mathbf{S}_h|\right], \text{ kde}$$

$$C_p = 1 + \frac{2p^2 + 3p - 1}{6(r-1)(p+1)}\left(\sum_{h=1}^{r}\frac{1}{n_h - 1} - \frac{1}{n-r}\right)$$ je konstanta zlepšující aproximaci.

V případě platnosti nulové hypotézy se statistika T_0 asymptoticky řídí rozložením $\chi^2\left(\frac{(r-1)p(p+1)}{2}\right)$. Pokud testová statistika nabude hodnoty aspoň $\chi^2_{1-\alpha}\left(\frac{(r-1)p(p+1)}{2}\right)$, hypotézu o shodě variančních matice zamítneme na asymptotické hladině významnosti α.

3.3. Test shody vektorů středních hodnot

Na hladině významnosti α testujeme nulovou hypotézu $H_0 : \boldsymbol{\mu}_1 = \ldots = \boldsymbol{\mu}_r$ proti alternativní hypotéze H_1 : aspoň jedna dvojice vektorů středních hodnot se liší.

Testy jsou založené na

- Wilksově kritériu,
- Lawleyově – Hotellingově kritériu,
- Pillaiově kritériu,
- Royově kritériu.

V praxi je nejpoužívanější Wilksovo kritérium $\Lambda = \frac{\det(\mathbf{E})}{\det(\mathbf{E}+\mathbf{B})}$. Nabývá hodnot mezi 0 a 1, přičemž vyšší hodnoty znamenají, že střední hodnoty se liší méně. Přitom $\mathbf{B}$ je matice meziskupinové variability a $\mathbf{E}$ je matice reziduální variability.

Testová statistika F_W pro test shody vektorů středních hodnot vznikne transformací Λ:

$F_W = -\left(n - \frac{p+r}{2} - 1\right)\ln\Lambda$. V případě platnosti nulové hypotézy se statistika F_W asymptoticky řídí rozložením $\chi^2(p(r-1))$. H_0 tedy zamítáme na asymptotické hladině významnosti α, když $F_W \geq \chi^2_{1-\alpha}(p(r-1))$.

4. LDA pro dvě skupiny objektů

4.1. Odvození bayesovského rozhodovacího pravidla

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

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

Nechť H_h je jev „objekt patří do h-té skupiny“.

Apriorní pravděpodobnost $P(H_h)$ příslušnosti objektu k h-té skupině označíme π_h, h = 1, 2.

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

$$P(H_h / \mathbf{X} = \mathbf{x}) = \frac{\pi_h \varphi_h(\mathbf{x})}{\pi_1 \varphi_1(\mathbf{x}) + \pi_2 \varphi_2(\mathbf{x})}, \quad h = 1, 2$$

Rozhodovací pravidlo: nový objekt zařadíme do té skupiny, u níž je aposteriorní pravděpodobnost větší.

Objekt s vektorem pozorování **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 $\pi_h\varphi_h(\mathbf{x})$ se nazývá diskriminační skór pro h-tou skupinu.

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

4.2. Odvození Fisherovy lineární diskriminační funkce pro dvě skupiny objektů

V diskriminační analýze se předpokládá, že hustota v h-té skupině je normální a má parametry $\boldsymbol{\mu}_h$, $\boldsymbol{\Sigma}_h$, tj.

$$\varphi_h(\mathbf{x}) = \frac{1}{\sqrt{\det(2\pi\boldsymbol{\Sigma}_h)}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_h)'\boldsymbol{\Sigma}_h^{-1}(\mathbf{x}-\boldsymbol{\mu}_h)\right), \text{ h = 1, 2.}$$

Jestliže zlogaritmujeme diskriminační skór $\pi_h\varphi_h(\mathbf{x})$ a vynecháme člen $-\frac{p}{2}\ln(2\pi)$, který je společný pro obě skupiny, dostaneme tzv. kvadratický diskriminační skór pro h-tou skupinu ve tvaru

$$-\frac{1}{2}\ln(\det\boldsymbol{\Sigma}_h) - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_h)'\boldsymbol{\Sigma}_h^{-1}(\mathbf{x}-\boldsymbol{\mu}_h) + \ln\pi_h, \text{ h = 1, 2.}$$

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 h-tou skupinu – tzv. Andersonovu diskriminační statistiku - ve tvaru $\lambda_h(\mathbf{x}) = \boldsymbol{\mu}_h'\boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_h'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_h + \ln\pi_h$, h = 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

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

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

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

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 **x**, která nám umožní minimalizovat celkovou pravděpodobnost mylného zařazení objektu do skupiny. Objekt s vektorem pozorování **x** tedy zařadíme do 1. skupiny, když $\lambda(\mathbf{x}) > 0$, jinak ho zařadíme do 2. skupiny.

4.3. 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:

$\boldsymbol{\mu}_h \rightarrow \mathbf{M}_h$, h = 1, 2

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

$$\pi_h \rightarrow \frac{n_h}{n}, \; h = 1, 2.$$

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

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

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

4.4. Posouzení účinnosti diskriminace resubstituční metodou

Resubstituční metoda spočívá v uplatnění zkonstruovaného rozhodovacího pravidla na objekty se známou příslušností ke skupině. Uvažujeme postupně všechny tyto objekty 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:

$$\frac{n_{11} + n_{22}}{n}$$

Podíl mylně zařazených objektů:

$$\frac{n_{12} + n_{21}}{n}$$

4.5. 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 a orientačně posoudíme linearitu vztahů mezi sledovanými proměnnými 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.

4.6. 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

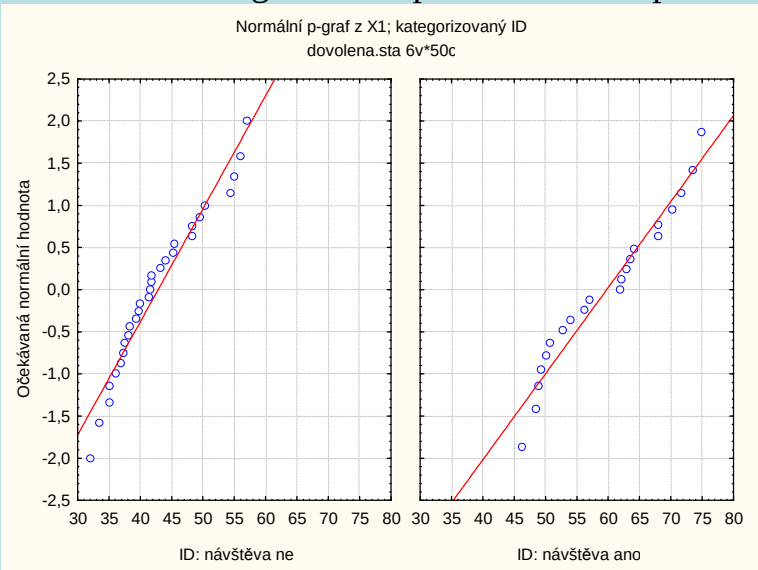
Proměnná	Testy normality (dovolena.sta) Zhrnout podmínku: ID=0		
	N	W	p
X1: roční příjem v tisících dolarů	29	0,940188	0,101411
X2: postoj k cestování (škála 9 bodů)	29	0,964071	0,412187
X3: význam rodinné dovolené (škála 9 bodů)	29	0,964432	0,420319
X4: počet členů rodiny	29	0,917696	0,026668
X5: věk nejstaršího člena	29	0,944508	0,131598

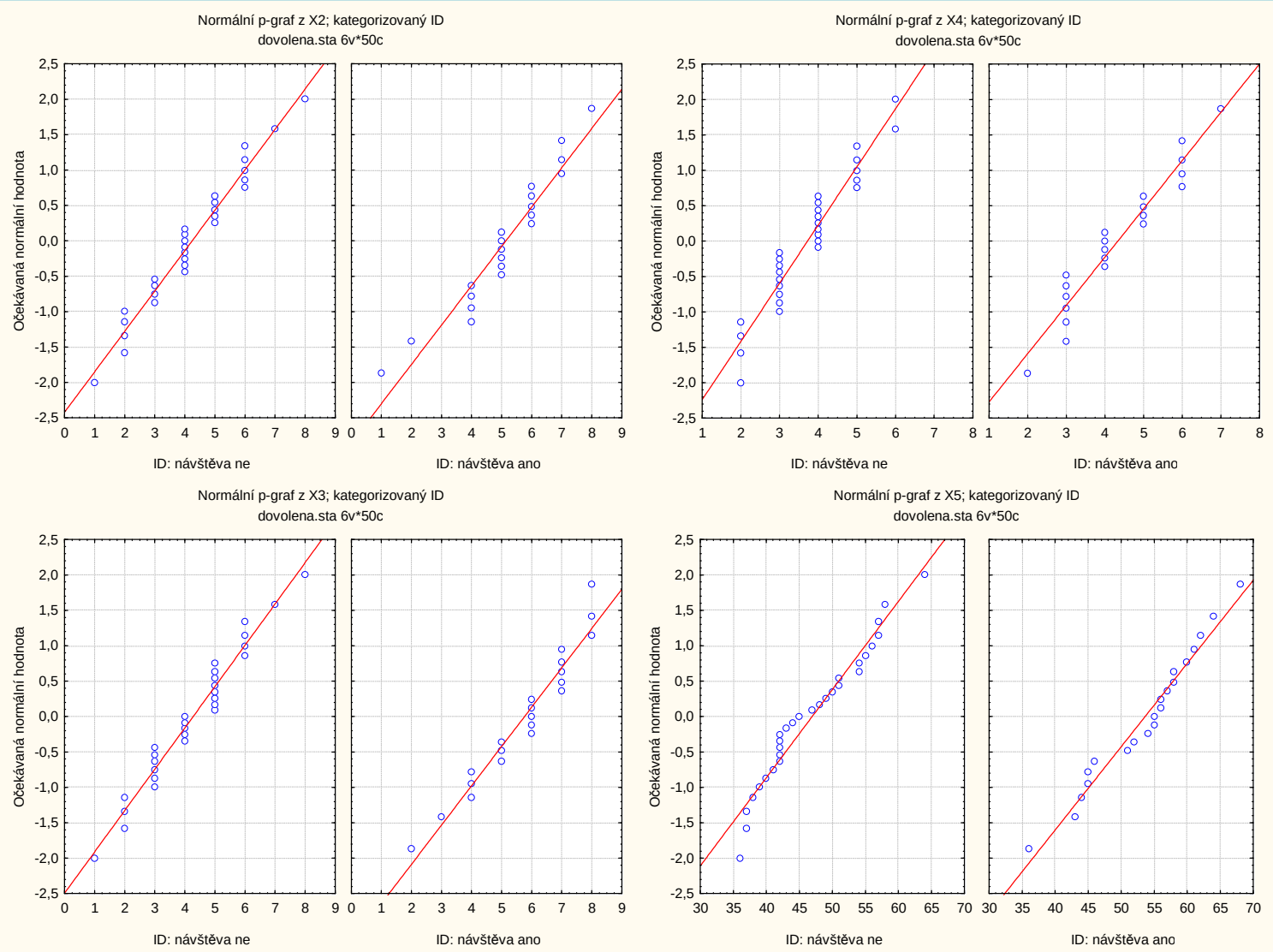
Pro skupinu rodin, které danou rekreační oblast navštěvují: Statistiky – Základní statistiky/tabulky – Select cases – ID=1 – OK – Tabulky četností – Proměnné X1 až X5 – OK – Normalita – zaškrtneme S-W test – Testy normality

Proměnná	Testy normality (dovolena.sta) Zhrnout podmínku: ID=1		
	N	W	p
X1: roční příjem v tisících dolarů	21	0,935874	0,180430
X2: postoj k cestování (škála 9 bodů)	21	0,930271	0,139382
X3: význam rodinné dovolené (škála 9 bodů)	21	0,934717	0,171087
X4: počet členů rodiny	21	0,928224	0,126815
X5: věk nejstaršího člena	21	0,967589	0,679311

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

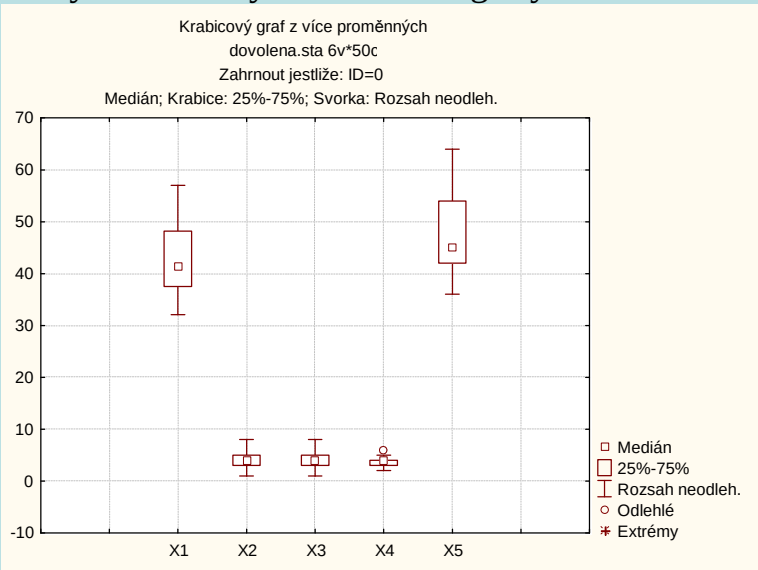



Normální p-graf z X2; kategorizovaný ID
dovolena.sta 6v*50c
Očekávaná normální hodnota
ID: návštěva ne
ID: návštěva ano
Normální p-graf z X4; kategorizovaný ID
dovolena.sta 6v*50c
Očekávaná normální hodnota
ID: návštěva ne
ID: návštěva ano
Normální p-graf z X3; kategorizovaný ID
dovolena.sta 6v*50c
Očekávaná normální hodnota
ID: návštěva ne
ID: návštěva ano
Normální p-graf z X5; kategorizovaný ID
dovolena.sta 6v*50c
Očekávaná normální hodnota
ID: návštěva ne
ID: návštěva ano

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

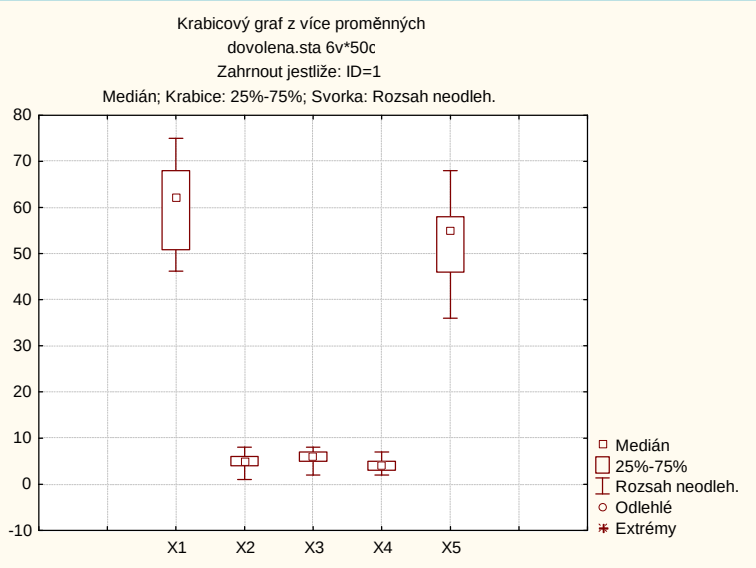
	Popisné statistiky (dovolena.sta) Zhrnout podmínku: ID=0	
Proměnná	N platných	Průměr
X1	29	42,84483
X2	29	4,24138
X3	29	4,27586
X4	29	3,72414
X5	29	46,93103

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



Nyní změníme podmínku ID = 1

Proměnná	Popisné statistiky (dovolena.sta) Zhrnout podmínku: ID=1 N platných	Průměr
X1	21	59,76190
X2	21	5,14286
X3	21	5,76190
X4	21	4,33333
X5	21	53,61905



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				
Proměnná	X1	X2	X3	X4	X5
X1	49,1947	0,99594	-2,24138	1,094951	-24,1647
X2	0,9959	2,76108	-0,31897	0,140394	-4,7328
X3	-2,2414	-0,31897	2,63547	-0,171182	1,1268
X4	1,0950	0,14039	-0,17118	1,278325	1,9446
X5	-24,1647	-4,73276	1,12685	1,944581	57,2808

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

	Kovariance (dovolena.sta) Zhrnout podmínku: ID=1				
Proměnná	X1	X2	X3	X4	X5
X1	83,59048	4,300714	6,39048	4,70333	16,25476
X2	4,30071	2,728571	0,03571	0,20000	1,05714
X3	6,39048	0,035714	2,79048	0,03333	-1,04524
X4	4,70333	0,200000	0,03333	1,83333	-2,46667
X5	16,25476	1,057143	-1,04524	-2,46667	63,84762

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	63,53	2,37	1,36	2,60	-7,32
X2	2,37	2,75	-0,17	0,17	-2,32
X3	1,36	-0,17	2,70	-0,09	0,22
X4	2,60	0,17	-0,09	1,51	0,11
X5	-7,32	-2,32	0,22	0,11	60,02

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: $W = \left\langle \chi^2_{1-\alpha}\left(\frac{p(p+1)}{2}\right), \infty \right) = \left\langle \chi^2_{0,95}(10), \infty \right) = \langle 24{,}9958, \infty \rangle$.

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

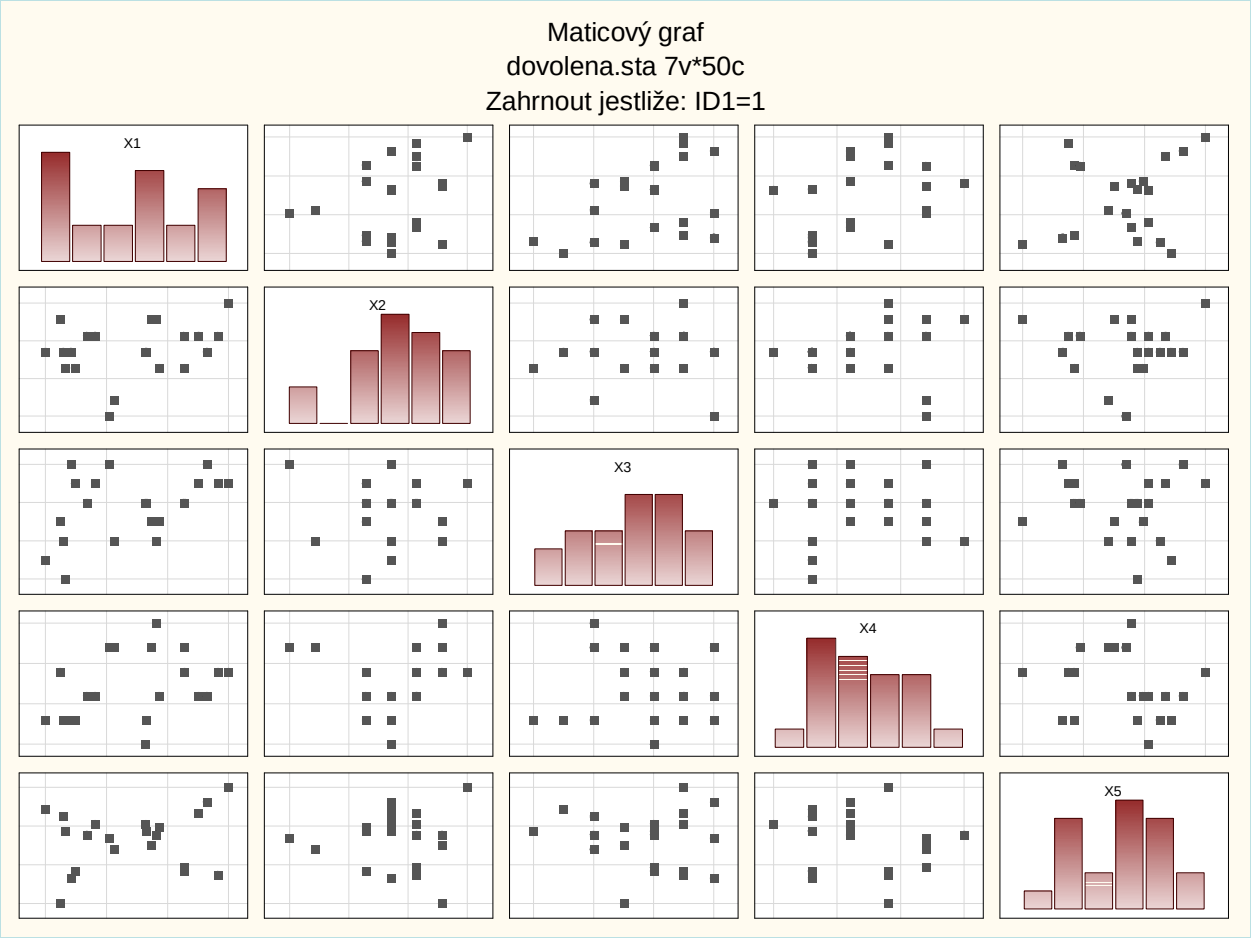
Provedení testu v systému STATISTICA:

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 – Boxův M-test.

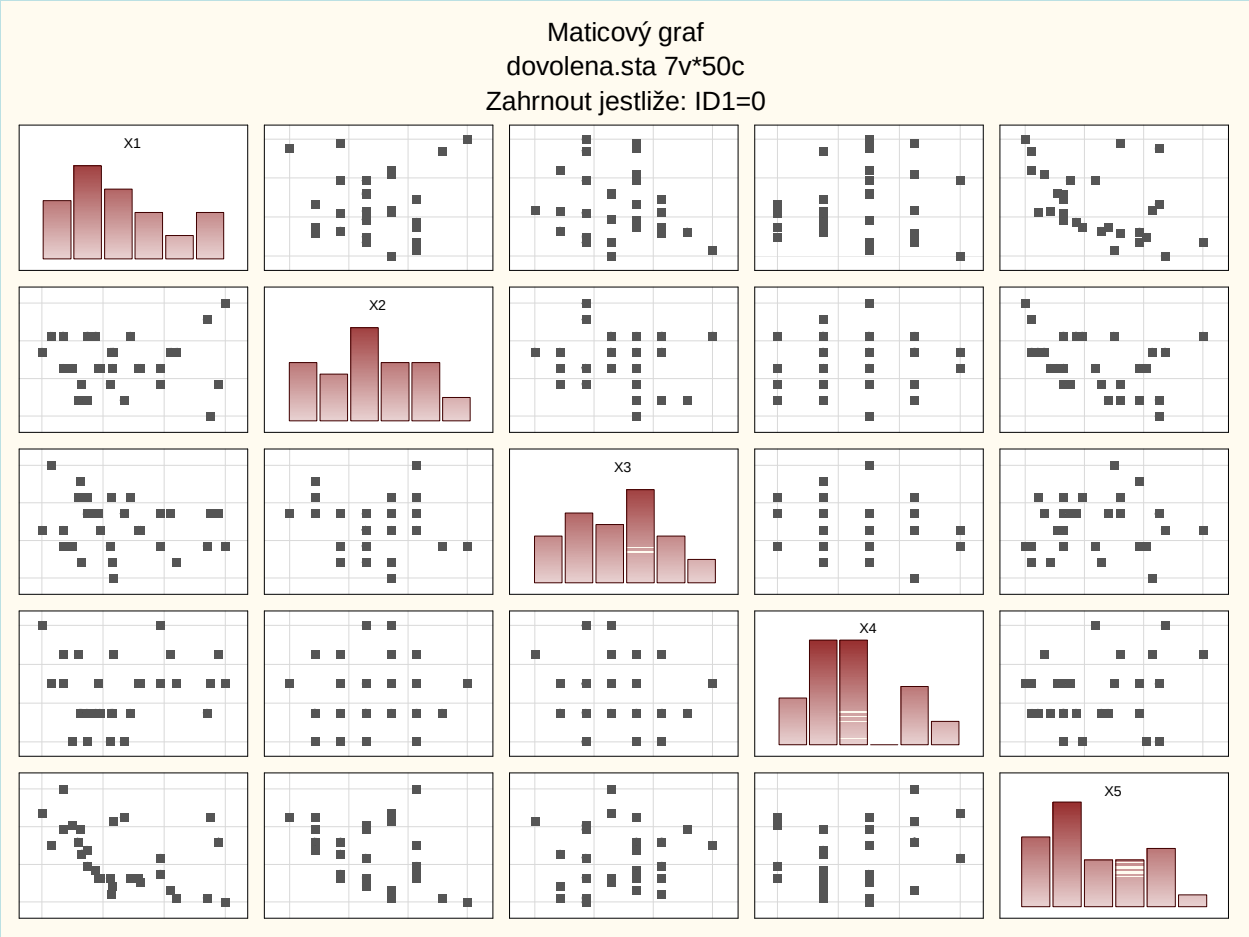
	Boxův M test (dovolena.sta) Efekt: ID (Vypočteno pro všechny proměnné)			
	Boxovo M	Chí-kv.	sv	p
Boxovo M	26,61690	23,54681	15	0,073200

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

Linearita vztahů mezi proměnnými ve skupině rodin navštěvujících danou oblast



Linearita vztahů mezi proměnnými ve skupině rodin nenavštěvujících danou oblast



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.

Proměnná	t-testy; grupováno: ID (dovolena.sta) Skup. 1: návštěva ne; Skup. 2: návštěva ano Hotellingovo 77,5606 F(5,44)=14,219 p<,00000								
	Průměr návštěva ne	Průměr návštěva ano	t	sv	p	Poč.plat návštěva ne	Poč.plat. návštěva ano	Sm.odch. návštěva ne	Sm.odch. návštěva ano
X1	42,84483	59,76190	-7,40751	48	0,000000	29	21	7,013894	9,142783
X2	4,24138	5,14286	-1,89805	48	0,063712	29	21	1,661651	1,651839
X3	4,27586	5,76190	-3,15623	48	0,002760	29	21	1,623412	1,670472
X4	3,72414	4,33333	-1,73042	48	0,089980	29	21	1,130630	1,354006
X5	46,93103	53,61905	-3,01289	48	0,004122	29	21	7,568407	7,990471

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.

Význam jednotlivých proměnných v modelu

Statistiky – Vícerozměrné průzkumné techniky – Diskriminační analýza – Proměnné - Grupovací ID1 – Seznam nezáv. proměnných X1 až X5 – OK – OK – Výpočet: proměnné v modelu.

N=50	Výsledky diskriminační funkční analýzy (dovolena.sta) Počet prom. v modelu: 5; grupovací: ID1 (2 skup) Wilk. lambda: ,38229 přibliž F (5,44)=14,219 p< ,0000					
	Wilk. Lambda	Parc. Lambda	F na vyj (1,44)	p-hodn.	Toler.	1-toler. R^2
X1	0,627513	0,609207	28,22504	0,000003	0,879866	0,120134
X2	0,388609	0,983729	0,72778	0,398223	0,934715	0,065285
X3	0,400086	0,955507	2,04884	0,159388	0,977164	0,022836
X4	0,382565	0,999270	0,03215	0,858527	0,921303	0,078697
X5	0,439319	0,870177	6,56444	0,013904	0,956782	0,043218

V záhlaví této tabulky je uvedena Wilksova Lambda (na škále od 0 – nejlepší diskriminace do 1 – žádná diskriminace) a její přepočet na testovou statistiku F pro Hotellingův test shody vektorů středních hodnot (14,219) a odpovídající p-hodnota (je blízká 0).
V 1. sloupci (Wilk. Lambda) jsou hodnoty Wilksovy Lambdy při vyřazení dané proměnné z modelu (vyšší hodnoty jsou lepší).
2. sloupec (Parc. Lambda) obsahuje unikátní příspěvky proměnných k diskriminaci.
Ve 3. sloupci jsou přepočty parciálních Lambda na testové statistiky a ve 4. sloupci pak odpovídající p-hodnoty. Podle p-hodnot u jednotlivých proměnných soudíme, že pro diskriminaci jsou významné proměnné X_1 a X_5.
5. sloupec (Tolerance) udává unikátní variabilitu proměnné nevysvětlenou ostatními proměnnými v modelu.
6. sloupec (1-toler., R^2) udává variabilitu proměnné vysvětlenou ostatními proměnnými.

Mahalanobisova vzdálenost v diskriminační analýze

Používá se pro popis vzájemných vzdáleností centroidů jednotlivých skupin.

Statistiky – Vícerozměrné průzkumné techniky – Diskriminační analýza - Proměnné – Grupovací proměnná ID, Seznam nezávislých proměnných X1 až X5 — OK – OK – na záložce Detaily zvolíme Vzdálenosti mezi skupinami. Současně dostaneme i p-hodnoty pro testy hypotéz, že vzdálenosti jsou nulové:

	Mahalanobisovy vzdálenosti^2 (dovolena.sta)	
ID1	návštěva ne	návštěva ano
návštěva ne	0,000000	6,367867
návštěva ano	6,367867	0,000000

	p-hodnot (dovolena.sta)	
ID1	návštěva ne	návštěva ano
návštěva ne		0,000000
návštěva ano	0,000000	

Lze také získat Mahalanobisovy vzdálenosti jednotlivých objektů od centroidů skupin.

Na záložce Klasifikace zvolíme Mahalanobisovy vzdálenosti^2:

	Mahalanobisovy vzdálenosti (dovolena.sta) Nesprávná klasifikace je označena *		
Případ	**Pozorova Klasif.**	návštěva ne p=,58000	návštěva ano p=,42000
1	návštěva ne	9,18363	18,11825
2	návštěva ne	0,88533	10,53314
3	návštěva ne	3,90372	12,30937
4	návštěva ne	5,35649	8,74744
5	návštěva ne	4,41397	11,30806
6	návštěva ne	0,62136	7,62423

Stanovení odhadu Fisherovy lineární diskriminační funkce:

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

Odhad vektoru středních hodnot v 1. skupině:

	Popisné statistiky (dovolena.sta) Zhrnout podmínku: ID=0	
Proměnná	N platných	Průměr
X1	29	42,84483
X2	29	4,24138
X3	29	4,27586
X4	29	3,72414
X5	29	46,93103

Odhad vektoru středních hodnot ve 2. skupině:

	Popisné statistiky (dovolena.sta) Zhrnout podmínku: ID=1	
Proměnná	N platných	Průměr
X1	21	59,76190
X2	21	5,14286
X3	21	5,76190
X4	21	4,33333
X5	21	53,61905

Odhad společné varianční matice **S**:

	X_1	X_2	X_3	X_4	X_5
X1	63,53	2,37	1,36	2,60	-7,32
X2	2,37	2,75	-0,17	0,17	-2,32
X3	1,36	-0,17	2,70	-0,09	0,22
X4	2,60	0,17	-0,09	1,51	0,11
X5	-7,32	-2,32	0,22	0,11	60,02

Postup v systému STATISTICA :

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.

Odhady apriorních pravděpodobností:

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

Po dosazení dostaneme:

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

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

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

Postup v systému STATISTICA :

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; grupovací : ID (dovolena
Proměnná	návštěva ne p=,58000	návštěva ano p=,42000
X1	0,6369	0,9054
X2	1,7840	2,0395
X3	1,3391	1,7560
X4	1,1866	1,1130
X5	0,9216	1,0743
Konstant	-44,6709	-69,4375

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

	Klasifikační funkce; grupovací : ID (dovolena		
Proměnná	návštěva ne p=,58000	návštěva ano p=,42000	NProm =v1-v2
X1	0,6369	0,9054	-0,26847
X2	1,7840	2,0395	-0,25557
X3	1,3391	1,7560	-0,41694
X4	1,1866	1,1130	0,073566
X5	0,9216	1,0743	-0,15266
Konstant	-44,6709	-69,4375	24,76658

Klasifikace nového případu

Předpokládejme nyní, že jsme prozkoumali další rodinu, která

má roční příjem $X_1 = 51{,}8$ tisíc dolarů,

k cestování zaujímá postoj ohodnocený $X_2 = 6$ body,

rodinné dovolené přičítá význam ohodnocený $X_3 = 7$ body,

má $X_4 = 4$ členy

a nejstaršímu členovi je $X_5 = 51$ 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.

Posouzení účinnosti diskriminace resubstituční metodou:

Na záložce Klasifikace zvolíme Klasifikační matice.

Skup.	Klasifikační matice (dovolena) Řádky: pozorované klasifikace Sloupce: předpovězené klasifikace % správných	návštěva ne p=,58000	návštěva ano p=,42000
návštěva ne	93,10345	27	2
návštěva ano	76,19048	5	16
Celkem	86,00000	32	18

Podíl správně zařazených objektů:

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

Podíl mylně zařazených objektů:

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

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.

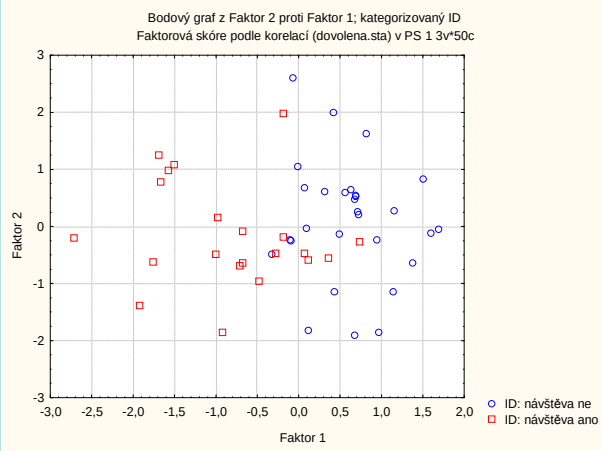
Porovnání s náhodnou klasifikací

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



5. Výběr proměnných pro klasifikaci krokovou metodou

Kroková metoda postupně vyhledává nejvhodnější soubor proměnných pro diskriminaci. Používá se buď jako dopředná nebo jako zpětná.

Význam jednotlivých proměnných pro diskriminaci se k každém kroku zkoumá pomocí zaváděcího a odstraňovacího kritéria.
Vybírání proměnných či jejich odstraňování skončí, když žádné další proměnné nesplňují zaváděcí nebo odstraňovací kritérium.

Upozornění: Před zařazením j-té proměnné do modelu se stanoví její tolerance $1-R_j^2$ (R_j^2 je čtverec vícenásobného koeficientu korelace, tj. koeficientu, který měří těsnost lineární závislosti veličiny X_j na ostatních veličinách). Tolerance je implicitně nastavená na 0,01.

Příklad: Použijte krokovou dopřednou (a poté zpětnou) metodu pro zařazování rodin do dvou skupin.

Řešení:

Statistika – Vícerozměrné průzkumné techniky – Diskriminační analýza – Proměnné - Grupovací ID1 – Seznam nezáv. proměnných X1 až X5 – OK – zaškrtneme Další možnosti (kroková analýza) – OK – Metoda – zvolíme kroková dopředná. Na záložce Detaily můžeme změnit Možnosti kroku (ponecháme implicitní nastavení) a také pomocí tlačítka Výsledky můžeme zvolit, zda chceme zobrazovat výsledky po každém kroku nebo chceme pouze shrnutí (ponecháme shrnutí) – OK.

Zvolíme-li tlačítko Výpočet: proměnné v modelu, dostaneme tabulku

N=50	Výsledky diskriminační funkční analýzy (dovolena.sta) krok 3, poč. prom. v modelu: 3; grupovací: ID1 (2 skup) Wilk. lambda: ,38880 přibliž F (3,46)=24,104 p< ,0000					
	Wilk. Lambda	Parc. Lambda	F na vyj (1,46)	p-hodn.	Toler.	1-toler. R^2
X1	0,719493	0,540386	39,12429	0,000000	0,974791	0,025209
X5	0,441811	0,880024	6,27128	0,015879	0,985042	0,014958
X3	0,405987	0,957678	2,03285	0,160683	0,988398	0,011602

Vidíme, že algoritmus skončil po třech krocích a vybral proměnné X_1, X_5 a X_3.

Zvolíme-li tlačítko Proměnné neobsažené v modelu, zjistíme, že jde o proměnné X_2 a X_4.
Na záložce Klasifikace vybereme Klasifikační funkce. Dostaneme lineární diskriminační skóry pro 1. a 2. skupinu objektů. Do vzniklé tabulky přidáme novou proměnnou L, do jejíhož Dlouhého jména napíšeme =v1-v2 a tím získáme odhad Fisherovy lineární diskriminační funkce:

	Klasifikační funkce; grupovací : ID1 (dovolena.sta)		
Proměnná	návštěva ne p=,58000	návštěva ano p=,42000	L =v1-v2
X1	0,7504	1,0247	-0,2742808
X5	0,8693	1,0128	-0,1434212
X3	1,1355	1,5365	-0,4009242
Konstant	-39,4479	-63,0649	23,6170025

Vidíme, že $L(\mathbf{x}) = -0{,}2743*X1 - 0{,}1434*X5 - 0{,}4009*X3 + 23{,}617$

Klasifikační matice je stejná jako v případě diskriminace podle všech proměnných a chybně zařazené případy jsou také stejné.

	Klasifikační matice (dovolena.sta) Řádky: pozorované klasifikace Sloupce: předpovězené klasifikace		
Skup.	% správnýc	návštěva ne p=,58000	návštěva ano p=,42000
návštěva ne	93,10345	27	2
návštěva ano	76,19048	5	16
Celkem	86,00000	32	18

Použijeme-li krokovou zpětnou metodu, je vybrána pouze proměnná X_1 a účinnost diskriminace poklesne na 80 %.

6. Lineární diskriminační analýza pro $r \geq 3$ skupin

6.1. Pravidlo pro zařazení objektu do skupiny

Opět předpokládáme, že ve všech r skupinách se vektory pozorování řídí p-rozměrným normálním rozložením, varianční matice jednotlivých skupin jsou shodné a vztahy mezi sledovanými p proměnnými jsou přibližně lineární.

Lineární diskriminační skór pro h-tou skupinu (Andersonova diskriminační statistika) má tvar:

$$\lambda_h(\mathbf{x}) = \boldsymbol{\mu}_h' \boldsymbol{\Sigma}^{-1} \mathbf{x} - \frac{1}{2} \boldsymbol{\mu}_h' \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_h + \ln \pi_h, \; h = 1, \ldots, r$$

Její odhad získáme dosazením $\mathbf{M}_h$, $\mathbf{S}$ a p_h:

$$L_h(\mathbf{x}) = \mathbf{M}_h' \mathbf{S}^{-1} \mathbf{x} - \frac{1}{2} \mathbf{M}_h' \mathbf{S}^{-1} \mathbf{M}_h + \ln p_h$$

Objekt neznámého původu, jehož vektor pozorování je $\mathbf{x}$, bude zařazen do skupiny s nejvyšší hodnotou $L_h(\mathbf{x})$.

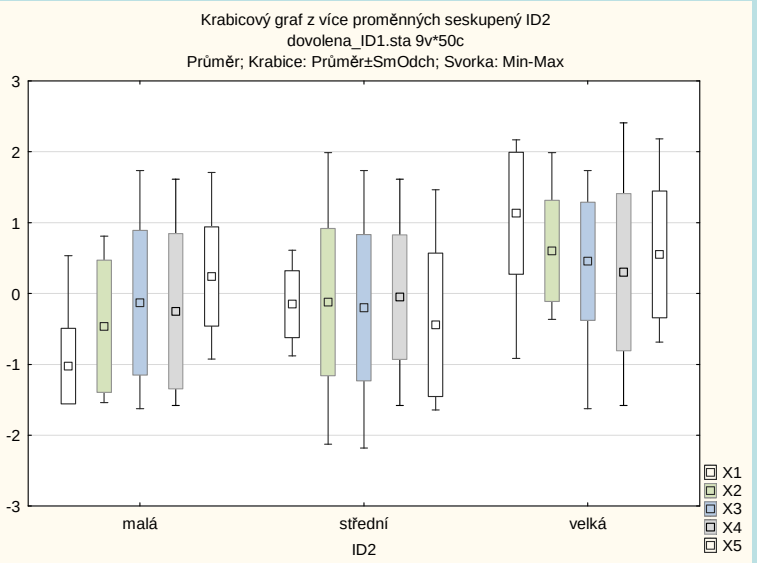
6.2. Příklad

Soubor rodin nyní roztřiďte do tří skupin podle proměnné ID2, tj. podle toho, jak velkou částku je rodina ochotna vydat z dovolenou (varianty „malá“, „střední“, „velká“).

Řešení:

Posouzení úrovně a variability proměnných X_1, …, X_5 v daných třech skupinách

Proměnná	ID2	N platných	Průměr	Sm.odch.
X1	malá	12	38,1	6,16
X2	malá	12	3,8	1,59
X3	malá	12	4,7	1,83
X4	malá	12	3,7	1,37
X5	malá	12	51,8	5,85
X1	střední	24	48,2	5,46
X2	střední	24	4,4	1,77
X3	střední	24	4,5	1,84
X4	střední	24	3,9	1,10
X5	střední	24	46,0	8,46
X1	velká	14	63,0	9,94
X2	velká	14	5,6	1,22
X3	velká	14	5,7	1,49
X4	velká	14	4,4	1,39
X5	velká	14	54,4	7,48



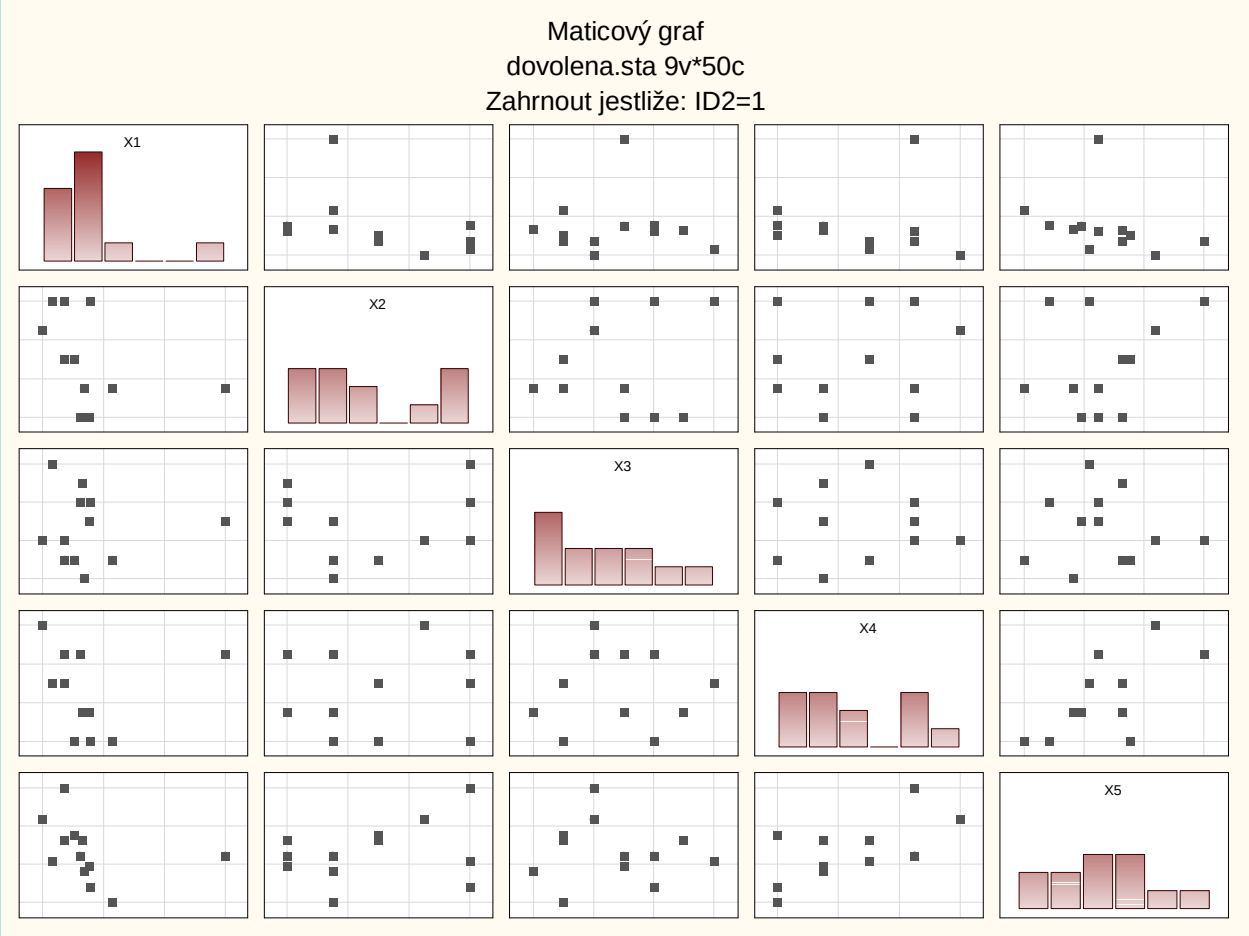
Ověření normality proměnných X_1, …, X_5 v daných třech skupinách

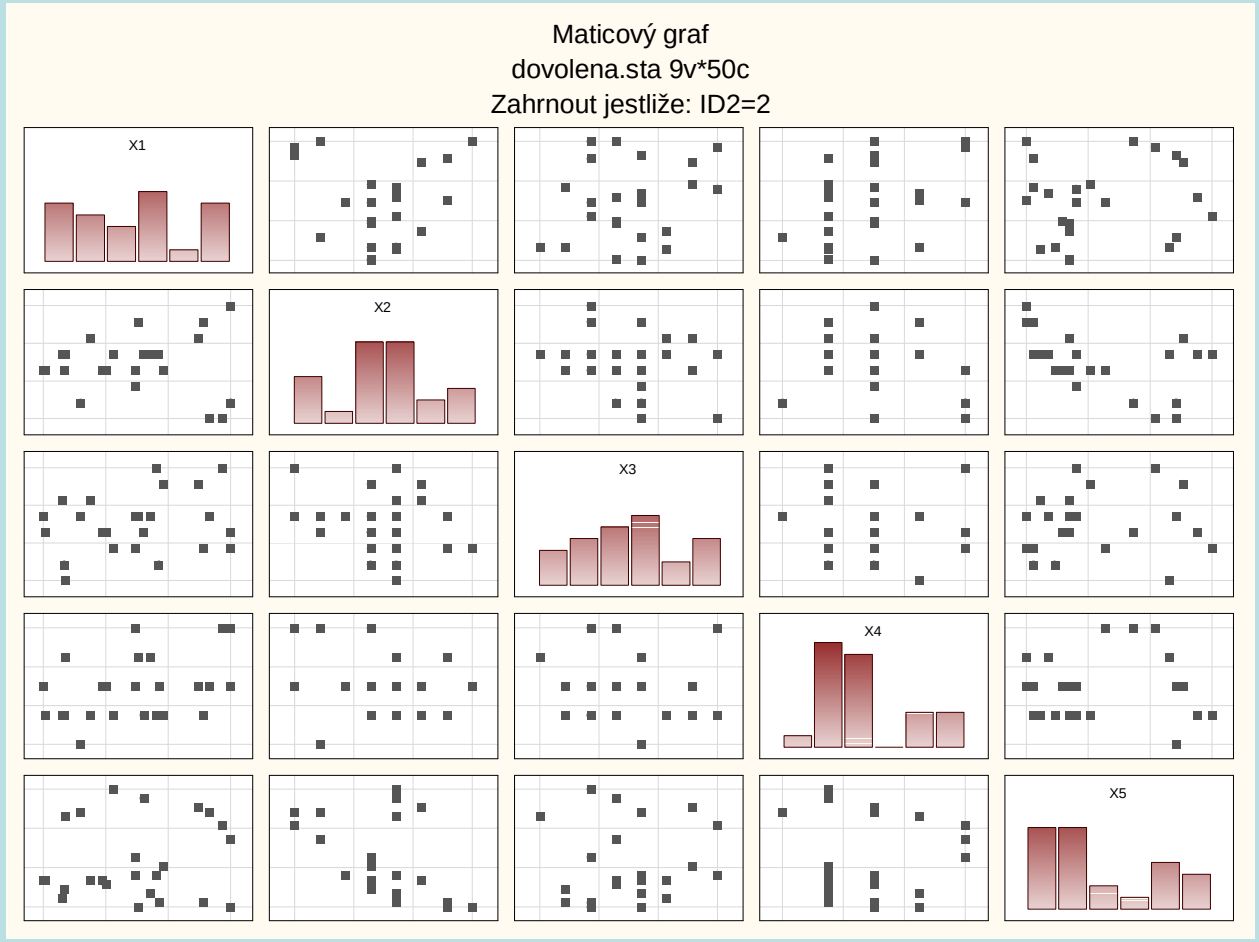
Proměnná	Souhrnné výsledky Testy normality (dovolena.sta) ID2	N	W	p
X1: roční příjem v tisících dolarů	malá	12	0,706875	0,000982
X2: postoj k cestování (škála 9 bodů)	malá	12	0,867375	0,060535
X3: význam rodinné dovolené (škála 9 bodů)	malá	12	0,955130	0,712720
X4: počet členů rodiny	malá	12	0,907871	0,200341
X5: věk nejstaršího člena	malá	12	0,976999	0,968796
X1: roční příjem v tisících dolarů	střední	24	0,947240	0,235912
X2: postoj k cestování (škála 9 bodů)	střední	24	0,943681	0,196939
X3: význam rodinné dovolené (škála 9 bodů)	střední	24	0,962008	0,480070
X4: počet členů rodiny	střední	24	0,877051	0,007252
X5: věk nejstaršího člena	střední	24	0,882154	0,009185
X1: roční příjem v tisících dolarů	velká	14	0,897737	0,104575
X2: postoj k cestování (škála 9 bodů)	velká	14	0,922488	0,238745
X3: význam rodinné dovolené (škála 9 bodů)	velká	14	0,909165	0,153244
X4: počet členů rodiny	velká	14	0,958259	0,694341
X5: věk nejstaršího člena	velká	14	0,933244	0,338619

Boxův test shody variančních matic

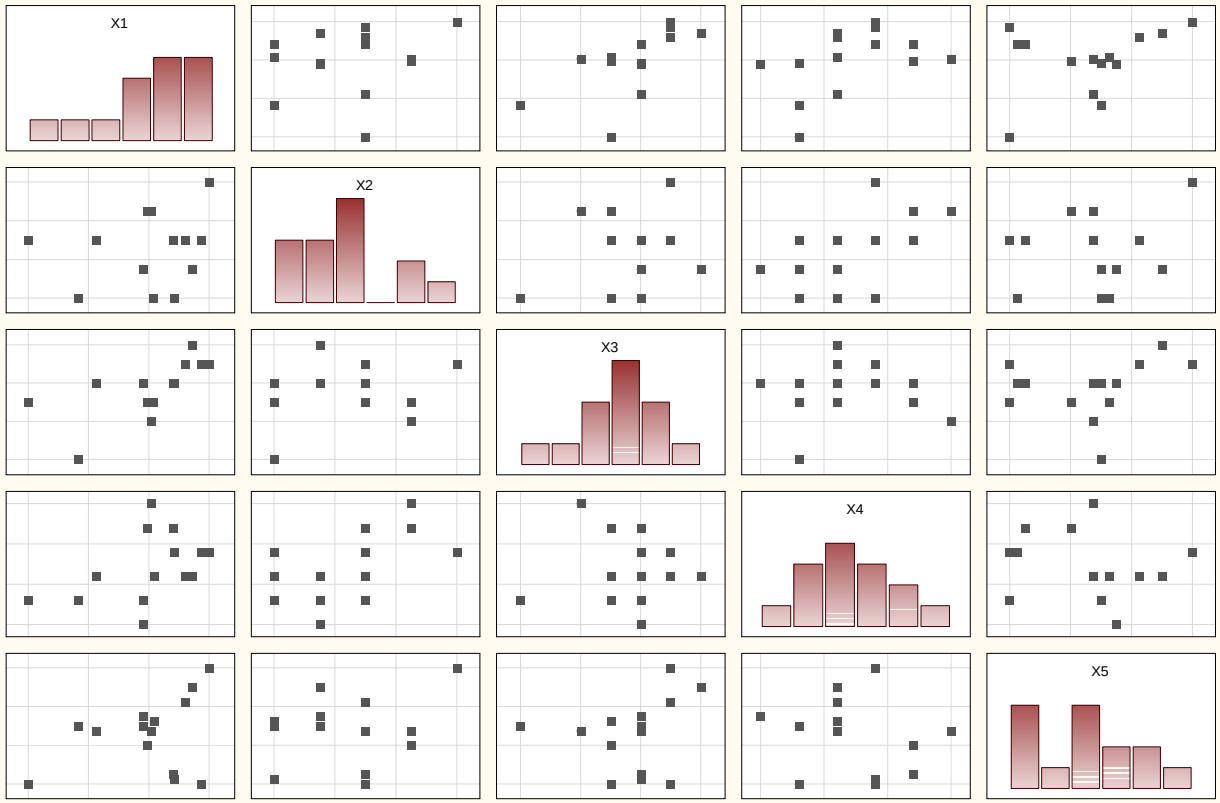
	Boxův M test (dovolena.sta) Efekt: "ID2" (Vypočteno pro všechny proměnné) Boxovo M	Chí-kv.	SV	p
Boxovo M	51,55790	42,84879	30	0,060418

Linearita vztahů proměnných X_1, …, X_5 v daných třech skupinách




Maticový graf
dovolena.sta 9v*50c
Zahrnout jestliže: ID2=2
X1
X2
X3
X4
X5

Maticový graf
dovolena.sta 9v*50c
Zahrnout jestliže: ID2=3



Testování hypotézy o shodě vektorů středních hodnot pomocí MANOVY

Efekt	Vícerozměrné testy významnosti. (dovolena.sta) Sigma-omezená parametrizace Dekompozice efektivní hypotézy Test	Hodnota	F	Efekt SV	Chyba SV	p
Abs. člen	Wilksův	0,01010	842,8765	5	43	0,000000
	Pillaiův	0,98990	842,8765	5	43	0,000000
	Hotellng	98,00890	842,8765	5	43	0,000000
	Royův	98,00890	842,8765	5	43	0,000000
"ID2"	Wilksův	0,26322	8,1626	10	86	0,000000
	Pillaiův	0,86784	6,7455	10	88	0,000000
	Hotellng	2,30122	9,6651	10	84	0,000000
	Royův	2,05945	18,1231	5	44	0,000000

Odlišnost vektorů středních hodnot ve sledovaných třech skupinách je prokázána na hladině významnosti 0,05.

Nyní provedeme simultánní testy o složkách vektorů středních hodnot.

Matice **E** reziduální variability

Efekt	Matice SSCP (Z' Z) reziduí (dovolena.sta) Sigma-omezená parametrizace Dekompozice efektivní hypotézy proměnné	X1	X2	X3	X4	X5
Chyba	X1	2386,662	-7,821	174,1762	134,0548	313,738
	X2	-7,821	118,714	-7,5119	5,9524	-103,131
	X3	174,176	-7,512	143,4821	1,1786	52,887
	X4	134,055	5,952	1,1786	73,7143	32,298
	X5	313,738	-103,131	52,8869	32,2976	2750,423

Matice **T** celkové variability

Efekt	Matice SSCP (Z' Z) odchylek (dovolena.sta) Matice SSCP (Z' Z) odchylek vektorů matice v matici schématu X Sloup.4 X1	Sloup.5 X2	Sloup.6 X3	Sloup.7 X4	Sloup.8 X5
X1	6535,025	299,6500	371,2500	250,2500	1026,550
X2	299,650	141,7800	8,1000	14,6200	-37,940
X3	371,250	8,1000	156,5000	6,9000	131,700
X4	250,250	14,6200	6,9000	76,9800	54,740
X5	1026,550	-37,9400	131,7000	54,7400	3425,620

Hodnoty testových statistik K1 až K5 a kritický obor:

	1 K1	2 K2	3 K3	4 K4	5 K5	6 kvantil
1	45,3276196	7,99016946	3,90805746	1,95069769	9,87874916	18,3070381

Na hladině významnosti 0,05 se prokázalo, že rozdíl mezi skupinami způsobuje X1.

Test shody vektorů středních hodnot a posouzení významu proměnných můžeme ve STATISTICE provést přímo v Diskriminační analýze.
Při zadávání proměnných zvolíme jako grupovací proměnnou ID2. Zvolíme-li Výpočet: proměnné v modelu, dostaneme tabulku:

N=50	Výsledky diskriminační funkční analýzy (dovolena.sta) Počet prom. v modelu: 5; grupovací: ID2 (3 skup) Wilk. lambda: ,26322 přibliž F (10,86)=8,1626 p< ,0000					
	Wilk. Lambda	Parc. Lambda	F na vyj (2,43)	p-hodn.	Toler.	1-toler. R^2
X1	0,602832	0,436636	27,74006	0,000000	0,805704	0,194297
X2	0,289522	0,909148	2,14852	0,129016	0,959666	0,040334
X3	0,270302	0,973794	0,57859	0,564991	0,899531	0,100469
X4	0,269947	0,975075	0,54960	0,581183	0,883696	0,116304
X5	0,319480	0,823896	4,59552	0,015533	0,948842	0,051158

V záhlaví této tabulky je uvedena testová statistika pro Wilksův test shody vektorů středních hodnot (8,1626) a odpovídající p-hodnota (je blízká 0).
Podle p-hodnot u jednotlivých proměnných soudíme, že pro diskriminaci jsou významné proměnné X_1 a X_5.

Na záložce Klasifikace zvolíme Klasifikační funkce:

Proměnná	Klasifikační funkce; grupovací : ID2 (dovolena.sta)		
	malá p=,24000	střední p=,48000	velká p=,28000
X1	0,5525	0,8026	1,0981
X2	2,3285	2,4727	3,1155
X3	0,6466	0,3530	0,3648
X4	0,7459	0,4926	0,1242
X5	0,8874	0,7754	0,9120
Konstant	-42,2581	-45,1663	-70,7708

Zde jsou uvedeny koeficienty pro odhady Andersonových diskriminačních skórů pro 1., 2. a 3. skupinu:

$L_1(\mathbf{x}) = 0{,}5525*X1 + 2{,}3285*X2 + 0{,}6466*X3 + 0{,}7459*X4 + 0{,}8874*X5 - 42{,}2581$
$L_2(\mathbf{x}) = 0{,}8026*X1 + 2{,}4727*X2 + 0{,}3530*X3 + 0{,}4926*X4 + 0{,}7754*X5 - 45{,}1663$
$L_3(\mathbf{x}) = 1{,}0981*X1 + 3{,}1155*X2 + 0{,}3648*X3 + 0{,}1242*X4 + 0{,}9120*X5 - 70{,}7708$

Klasifikační matice:

Skup.	Klasifikační matice (dovolena.sta) Řádky: pozorované klasifikace Sloupce: předpovězené klasifikace % správnýc	malá p=,24000	střední p=,48000	velká p=,28000
malá	66,66666	8	4	0
střední	91,66666	1	22	1
velká	78,57143	0	3	11
Celkem	82,00000	9	29	12

Správně zařazeno bylo $\frac{8+22+11}{50}\cdot 100\% = 82\%$ případů, chybně 18 % případů.

V 1. skupině rodin byly chybně zařazeny případy 8, 10, 19, 20 ($\frac{4}{12} = 33{,}3\%$), ve 2. skupině případy 4, 47 ($\frac{2}{24} = 8{,}3\%$) a ve 3. skupině případy 24, 34, 43 ($\frac{3}{14} = 21{,}4\%$)

Zařazení nového případu

Nyní podle těchto skórů zařadíme do jedné ze tří skupin rodinu, která

má roční příjem X_1 = 51,8 tisíc dolarů,

k cestování zaujímá postoj ohodnocený X_2 = 6 body,

rodinné dovolené přičítá význam ohodnocený X_3 = 7 body,

má X_4 = 4 členy

a nejstaršímu členovi je X_5 = 51 let.

Otevřeme nový datový soubor s osmi proměnnými a jedním případem. Do prvních pěti proměnných napíšeme zadané hodnoty a do Dlouhých jmen posledních tří proměnných napíšeme vyjádření pro odhady diskriminačních skórů.

	1 X1	2 X2	3 X3	4 X4	5 X5	6 L1	7 L2	8 L3
1	51,8	6	7	4	51	53,0996	55,23138	54,36618

Největší hodnotu má skór ve 2. skupině, tedy zkoumaná rodina vydá za dovolenou střední částku.

Dále v LDA použijeme pro výběr proměnných krokovou metodu.

Výsledky pro krokovou dopřednou metodu

Proměnné obsažené v modelu

N=50	Wilk. Lambda	Parc. Lambda	F na vyj (2,45)	p-hodn.	Toler.	1-toler. R^2
	Výsledky diskriminační funkční analýzy (dovolena.sta) krok 3, poč. prom. v modelu: 3; grupovací: ID2 (3 skup) Wilk. lambda: ,27663 přibliž F (6,90)=13,519 p< ,0000					
X1	0,652311	0,424084	30,55552	0,000000	0,984948	0,015052
X5	0,338537	0,817147	5,03482	0,010635	0,953070	0,046930
X2	0,303098	0,912692	2,15236	0,128024	0,967370	0,032630

Klasifikační funkce

Proměnná	malá p=,24000	střední p=,48000	velká p=,28000
	Klasifikační funkce; grupovací : ID2 (dovolena.sta)		
X1	0,6401	0,8551	1,1311
X5	0,8991	0,7824	0,9163
X2	2,3409	2,4846	3,1046
Konstant	-41,3768	-44,8553	-70,5840

Klasifikační matice

Skup.	% správnýc	malá p=,24000	střední p=,48000	velká p=,28000
	Klasifikační matice (dovolena.sta) Řádky: pozorované klasifikace Sloupce: předpovězené klasifikace			
malá	75,00000	9	3	0
střední	83,33334	3	20	1
velká	78,57143	0	3	11
Celkem	80,00000	12	26	12

Úspěšnost klasifikace poklesla z 82 % na 80 %.

Výsledky pro krokovou zpětnou metodu

Proměnné obsažené v modelu

	Výsledky diskriminační funkční analýzy (dovolena.sta) krok 4, poč. prom. v modelu: 1; grupovací: ID2 (3 skup) Wilk. lambda: ,36521 přibliž F (2,47)=40,846 p< ,0000					
N=50	Wilk. Lambda	Parc. Lambda	F na vyj (2,47)	p-hodn.	Toler.	1-toler. R^2
X1	1,000000	0,365211	40,84639	0,000000	1,000000	0,00

Klasifikační funkce

	Klasifikační funkce; grupovací : ID2 (dovolena.sta)		
Proměnná	malá p=,24000	střední p=,48000	velká p=,28000
X1	0,7506	0,9498	1,2413
Konstant	-15,7327	-23,6411	-40,3976

Klasifikační matice

	Klasifikační matice (dovolena.sta) Řádky: pozorované klasifikace Sloupce: předpovězené klasifikace			
Skup.	% správnýc	malá p=,24000	střední p=,48000	velká p=,28000
malá	83,3333	10	2	0
střední	100,0000	0	24	0
velká	78,5714	1	2	11
Celkem	90,0000	11	28	11

Je-li ke klasifikaci rodin do skupin použita pouze proměnná X_1, je úspěšnost klasifikace nejvyšší, a to 90 %.
Aplikujeme-li toto klasifikační pravidlo na rodinu s vektorem pozorování (51,8 6 7 4 51)', dostaneme výsledek

	1 X1	2 X2	3 X3	4 X4	5 X5	6 L1	7 L2	8 L3
1	51,8	6	7	4	51	23,14838	25,55854	23,90174