Diskriminační analýza

Otázky:

- Je možné předem stanovené skupiny objektů odlišit na základě proměnných, které máme zjištěné pro každý objekt?
- Které proměnné přispívají k tomuto odlišení největší měrou?
- 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.)

Na první dvě otázky odpovídá kanonická diskriminační analýza, na třetí pak klasifikační diskriminační analýza.

Možnosti použití diskriminační analýzy

1. 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.

2. 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.

3. 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.

4. 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.

Kanonická diskriminační analýza

Kanonická diskriminační analýza je metoda, která umožňuje sledovat vztahy mezi objekty v tzv. kanonickém prostoru, tj. prostoru vymezeném kanonickými proměnnými. Lze ji však využít i pro klasifikaci objektů s neznámou příslušností.

Předpokládáme, že máme $r \geq 2$ skupin objektů, v h-té skupině je n_h objektů a i-tý objekt je popsán p proměnnými $X_1, \ldots, X_p$. Tedy X_{hij} je hodnota j-té proměnné na i-tém objektu, který patří do h-té skupiny, $h = 1,\ldots,r,\ i = 1,\ldots,n_h,\ j = 1,\ldots,p$. Přitom $n_h > p$. Všech objektů je $n = \sum_{h=1}^{r} n_h$.

Hypotetický objekt z h-té skupiny, jehož vektor pozorování je stejný jako vektor výběrových průměrů v h-té skupině, se nazývá centroid h-té skupiny.

Mahalanobisova vzdálenost objektu s vektorem pozorování $\mathbf{x} = (x_1,\ldots,x_p)^T$ od centroidu h-té skupiny je dána vzorcem: $M(\mathbf{x},\mathbf{m_h}) = (\mathbf{x}-\mathbf{m_h})^T \mathbf{S_h}^{-1} (\mathbf{x}-\mathbf{m_h})$, kde $\mathbf{S_h}$ je varianční matice h-té skupiny a $\mathbf{m}_h$ je vektor průměrů h-té skupiny.

Celková variabilita obsažená v datech je vyjádřena maticí **T**, která se rozpadá na matici vnitroskupinové (reziduální) variability **E** a matici meziskupinové variability **B**: **T** = **E** + **B**.

Kanonická diskriminační analýza hledá takovou lineární kombinaci

$$Y = a_1 X_1 + \ldots + a_p X_p = a^T X$$

daných p proměnných, aby podíl meziskupinové a vnitroskupinové variability byl co největší.

Podíl $\lambda = \frac{\mathbf{a}^T \mathbf{B} \mathbf{a}}{\mathbf{a}^T \mathbf{E} \mathbf{a}}$ se nazývá Fisherovo diskriminační kritérium a hledáme jeho maximum vzhledem k **a**.

Vektor **a** získáme tak, že funkci $\lambda = \frac{\mathbf{a}^T \mathbf{B} \mathbf{a}}{\mathbf{a}^T \mathbf{E} \mathbf{a}}$ parciálně derivujeme podle $a_1, \ldots, a_p$, derivace položíme rovny 0 a řešíme systém p rovnic pro p neznámých. Tento systém lze maticově zapsat ve tvaru $\mathbf{B}\mathbf{E}^{-1}\mathbf{a} = \lambda \mathbf{a}$. Z lineární algebry je známo, že tento systém má netriviální řešení, právě když charakteristický polynom matice $\mathbf{B}\mathbf{E}^{-1}$ je nulový, tj. platí charakteristická rovnice $\left|\mathbf{B}\mathbf{E}^{-1} - \lambda \mathbf{I}\right| = 0$.

Řešením získáme k vlastních čísel $\lambda_1 > \ldots > \lambda_k$ matice $\mathbf{B}\mathbf{E}^{-1}$, kde $k = \min\{p, r-1\}$.

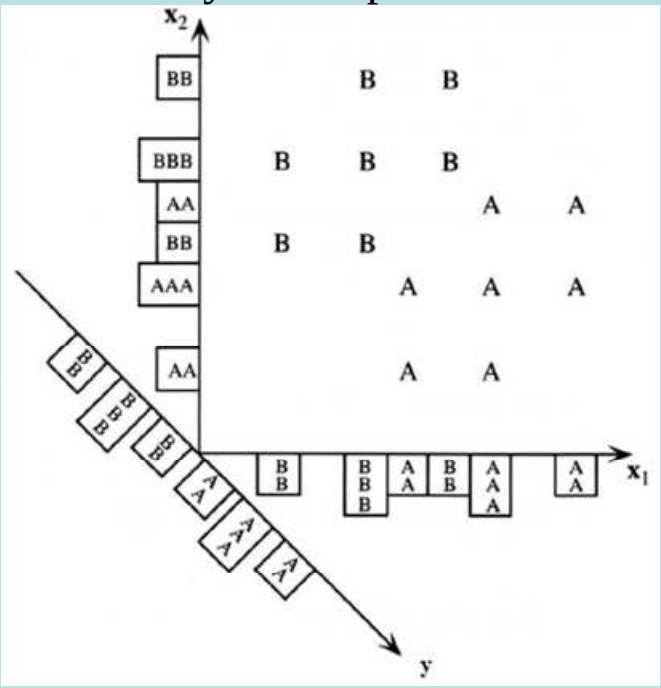
K nim jsou příslušné vlastní vektory $\mathbf{a}_1, \ldots, \mathbf{a}_k$.

Největšímu vlastnímu číslu λ_1 odpovídá vlastní vektor $\mathbf{a}_1$, který maximalizuje Fisherovo diskriminační kritérium.

Poznámka: Charakteristická rovnice neurčuje vektor $\mathbf{a}_1$ jednoznačně, ale pouze stanovuje poměr mezi jeho složkami. Konkrétní hodnoty složek vektoru $\mathbf{a}_1$ lze určit např. tak, aby platilo: $\mathbf{a_1}^{\mathrm{T}}\mathbf{a_1} = 1$, tedy aby vektor $\mathbf{a}_1$ by normovaný. Výhodnější je však volit $\mathbf{a}_1$ tak, aby $\frac{1}{n-r}\mathbf{a_1}^{\mathrm{T}}\mathbf{E}\mathbf{a_1} = 1$. Pak charakteristické číslo $\lambda_1 = \mathbf{a_1}^{\mathrm{T}}\mathbf{B}\mathbf{a_1}$ představuje míru meziskupinové variability veličiny Y_1 a složkám vektoru $\mathbf{a}_1$ se říká standardizované koeficienty.

Lineární kombinace $Y_1 = \mathbf{a_1}^{\mathrm{T}}\mathbf{X}$ se nazývá 1. kanonická proměnná (nebo též 1. diskriminant, 1. kanonická funkce). Geometricky lze Y_1 chápat jako projekci bodů reprezentujících jednotlivé objekty v p-rozměrném prostoru na přímku (tzv. diskriminační přímku), která umožňuje největší diskriminaci mezi centroidy skupin objektů.

Ilustrace významu první kanonické proměnné (p = 2):



Diskriminační přímka je vedena ve směru největší variability mezi skupinami.

Jsou-li objekty rozděleny do dvou skupin, stačí použít 1. kanonickou proměnnou. Jsou-li objekty roztříděny do více než dvou skupin, musíme použít další kanonické proměnné $Y_l = \mathbf{a}_l^T \mathbf{X}$, $l = 2, \ldots, k$. Kanonické proměnné jsou nezávislé, jsou uspořádány podle klesajícího významu a vymezují kanonický prostor.

Podíl $\frac{\lambda_l}{\lambda_1 + \ldots + \lambda_k}$ podává informaci o tom, jak se l-tá kanonická proměnná podílí na odlišení jednotlivých skupin, $l = 1, \ldots, k$.

Standardizovaný koeficient a_{lj} lze interpretovat jako vliv j-té proměnné na l-tou kanonickou proměnnou za předpokladu, že ostatních p-1 původních proměnných je konstantní.

Současně posuzujeme koeficienty korelace původních proměnných s kanonickými proměnnými. Vysoká absolutní hodnota koeficientu korelace některé proměnné s kanonickou proměnnou totiž znamená, že tato proměnná je pro kanonickou proměnnou charakteristická.

Zařazování objektů do skupin

Uvažme l-tou kanonickou proměnnou $Y_l = \mathbf{a}_l^T \mathbf{X}$ a i-tý objekt v h-té skupině s vektorem pozorování $\mathbf{x}_{hi} = (x_{hi1}, \ldots, x_{hip})^T$. Výraz $y_{hli} = c_l + \sum_{j=1}^{p} a_{lj} x_{hij}$, kde $c_l = -\mathbf{a}_l^T \mathbf{m}$ (**m** je vektor výběrových průměrů daných p-proměnných) se nazývá l-té diskriminační skóre i-tého objektu v h-té skupině.

Průměrné hodnoty jednotlivých kanonických proměnných ve skupinách se nazývají skupinové centroidy kanonických proměnných a jsou dány vzorcem: $\bar{y}_{hl} = c_l + \sum_{j=1}^{p} a_{lj} m_{hj}$.

Uvažme prvních s kanonických proměnných. Máme objekt s neznámou příslušností ke skupině, přičemž jeho vektor pozorování je **x**. Označme y_l jeho l-té diskriminační skóre, l = 1, …, s. Vypočítáme vzdálenost tohoto objektu od h-tého skupinového centroidu kanonických proměnných, h = 1, …, r:

$d_h^2 = \sum_{l=1}^{s} (y_l - \bar{y}_{hl})^2$. Objekt zařadíme do té skupiny, pro kterou bude vzdálenost d_h^2 nejmenší.

Upozornění 1: Vzdálenost d_h^2 je vlastně kvadrát Mahalanobisovy vzdálenosti objektu v kanonickém prostoru od h-tého skupinového centroidu kanonických proměnných.

Upozornění 2: Toto pravidlo zařazování nebere do úvahy velikost skupin.

Klasifikační matice

Úspěšnost zařazování objektů do skupin můžeme posoudit tak, že aplikujeme výše popsané zařazovací pravidlo na každý objekt a zařazení objektů porovnáme s jejich skutečnou příslušností ke skupině. Stanovíme podíl správně a mylně zařazených objektů.

skutečnost	zařazení			součet
	1. skupina	…	r-tá skupina	
1. skupina	n_{11}	…	n_{1r}	$n_{1.} = n_1$
…	…	…	…	
r-tá skupina	n_{r1}	…	n_{rr}	$n_{r.} = n_r$
součet	$n_{.1}$	…	$n_{.r}$	n

Podíl správně zařazených objektů: $\frac{n_{11} + \ldots + n_{rr}}{n}$

Podíl mylně zařazených objektů: $1 - \frac{n_{11} + \ldots + n_{rr}}{n}$

Příklad:

V souboru 50 rodin byly zjišťovány tyto údaje:

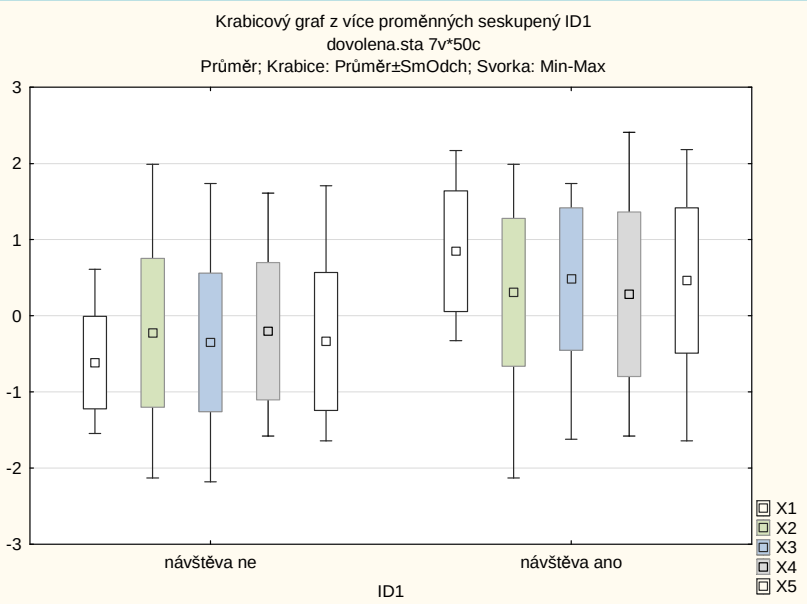
- zda v posledních dvou letech rodina navštívila jistou rekreační oblast (veličina ID1, nabývá hodnoty 0 pro odpověď „ne“, hodnoty 1 pro odpověď „ano“)
- částka, kterou je rodina ochotná vydat za dovolenou (veličina ID2, nabývá hodnoty 1 pro variantu „malá“, 2 pro variantu „střední“ a 3 pro variantu „velká“)
- 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 proveďte kanonickou diskriminační analýzu, a to pro třídění jak pomocí proměnné ID1 (dvě skupiny), tak pomocí proměnné ID2 (tři skupiny).

Řešení pro dvě skupiny (třídění podle ID1):

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

Proměnná	Souhrnné výsledky Popisné statistiky (dovolena.sta) ID1	N platných	Průměr	Sm.odch.
X1	návštěva ne	29	42,8	7,014
X2	návštěva ne	29	4,2	1,662
X3	návštěva ne	29	4,3	1,623
X4	návštěva ne	29	3,7	1,131
X5	návštěva ne	29	46,9	7,568
X1	návštěva ano	21	59,8	9,143
X2	návštěva ano	21	5,1	1,652
X3	návštěva ano	21	5,8	1,670
X4	návštěva ano	21	4,3	1,354
X5	návštěva ano	21	53,6	7,990



Ověření normality proměnných $X_1, \ldots, X_5$ v daných dvou skupinách

Proměnná	Souhrnné výsledky Testy normality (dovolena.sta) ID1	N	W	p
X1: roční příjem v tisících dolarů	návštěva ne	29	0,940188	0,101411
X2: postoj k cestování (škála 9 bodů)	návštěva ne	29	0,964071	0,412187
X3: význam rodinné dovolené (škála 9 bodů)	návštěva ne	29	0,964432	0,420319
X4: počet členů rodiny	návštěva ne	29	0,917696	0,026668
X5: věk nejstaršího člena	návštěva ne	29	0,944508	0,131598
X1: roční příjem v tisících dolarů	návštěva anc	21	0,935874	0,180430
X2: postoj k cestování (škála 9 bodů)	návštěva anc	21	0,930271	0,139382
X3: význam rodinné dovolené (škála 9 bodů)	návštěva anc	21	0,934717	0,171087
X4: počet členů rodiny	návštěva anc	21	0,928224	0,126815
X5: věk nejstaršího člena	návštěva anc	21	0,967589	0,679311

Odhad varianční matice $\mathbf{S}_1$

Proměnná	Kovariance (dovolena.sta) Zhrnout podmínku: ID1=0 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

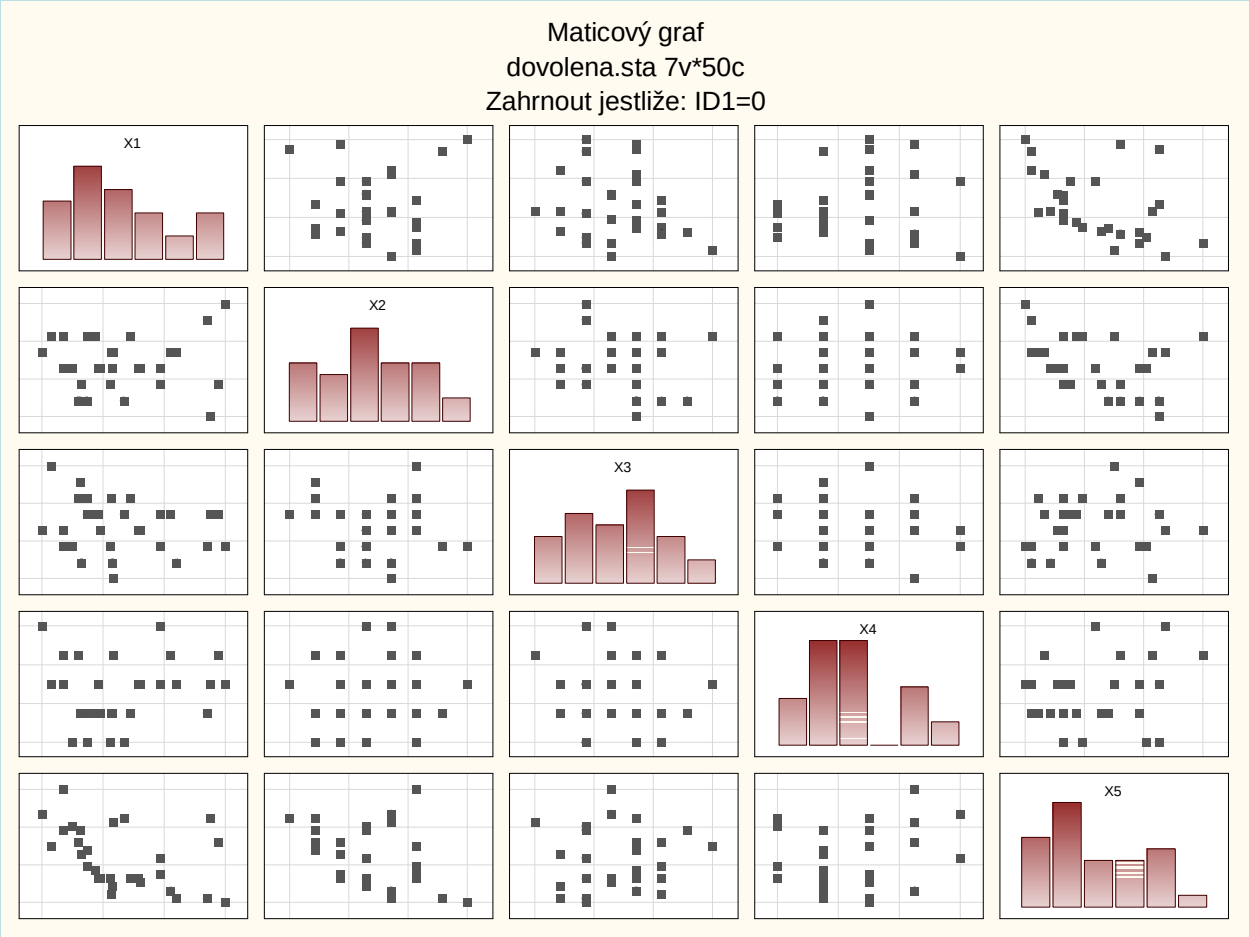
Odhad varianční matice $\mathbf{S}_2$

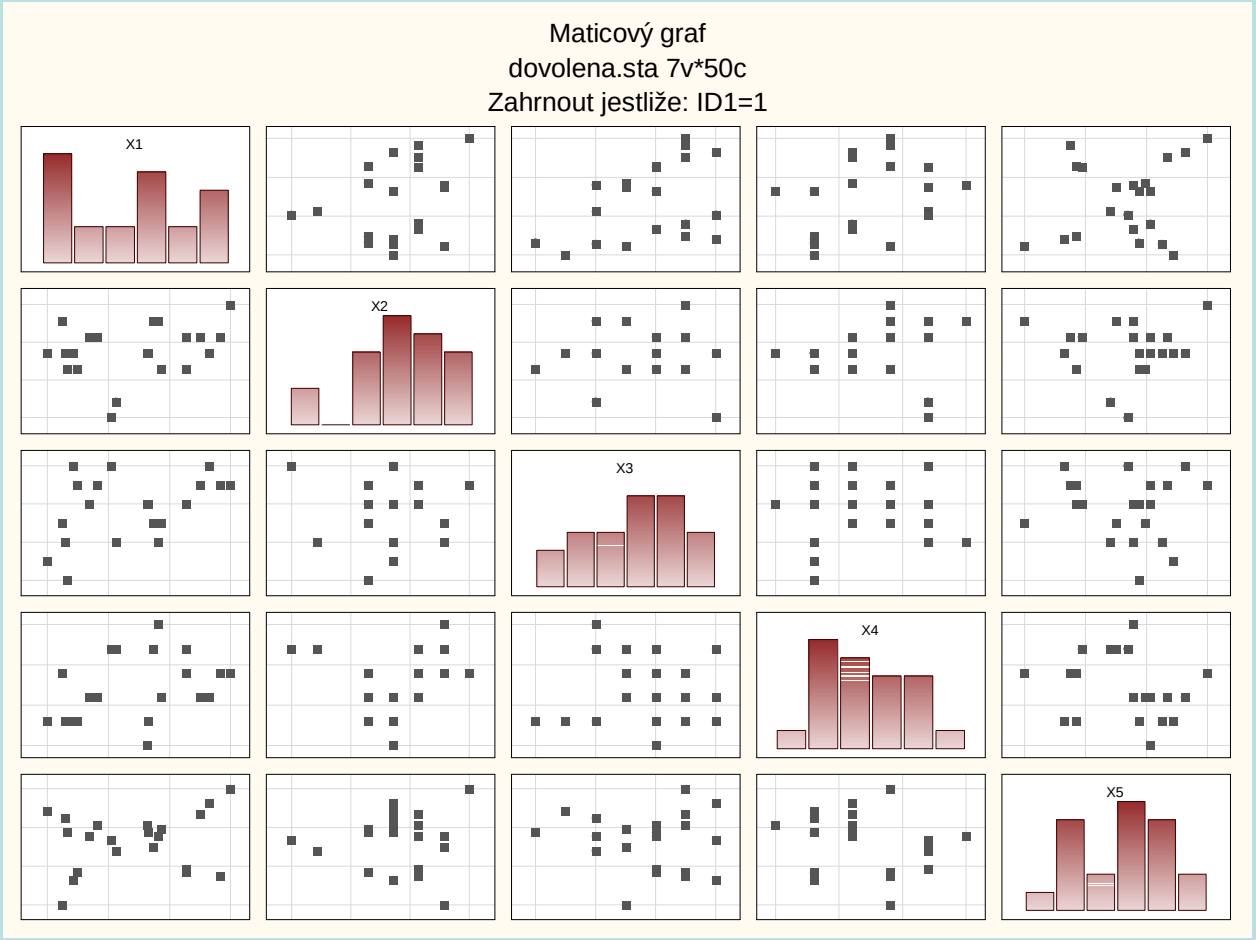
Proměnná	Kovariance (dovolena.sta) Zhrnout podmínku: ID1=1 X1	X2	X3	X4	X5
X1	83,59048	4,300714	6,39048	4,70333	16,25476
X2	4,30071	2,728571	0,03571	0,20000	1,05714
X3	6,39048	0,035714	2,79048	0,03333	-1,04524
X4	4,70333	0,200000	0,03333	1,83333	-2,46667
X5	16,25476	1,057143	-1,04524	-2,46667	63,84762

Boxův test shody variančních matic

	Boxův M test (dovolena.sta) Efekt: "ID1" (Vypočteno pro všechny proměnné) Boxovo M	Chí-kv.	SV	p
Boxovo M	26,61690	23,54681	15	0,073200

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




Maticový graf
dovolena.sta 7v*50c
Zahrnout jestliže: ID1=1
X1
X2
X3
X4
X5

Odhad korelační matice $\mathbf{R}_1$

Proměnná	Korelace (dovolena.sta) Zhrnout podmínku: ID1=0				
	X1	X2	X3	X4	X5
X1	1,000000	0,085454	-0,196846	0,138075	-0,455215
X2	0,085454	1,000000	-0,118243	0,074729	-0,376331
X3	-0,196846	-0,118243	1,000000	-0,093263	0,091713
X4	0,138075	0,074729	-0,093263	1,000000	0,227248
X5	-0,455215	-0,376331	0,091713	0,227248	1,000000

Odhad korelační matice $\mathbf{R}_2$

Proměnná	Korelace (dovolena.sta) Zhrnout podmínku: ID1=1				
	X1	X2	X3	X4	X5
X1	1,000000	0,284770	0,418423	0,379933	0,222500
X2	0,284770	1,000000	0,012943	0,089421	0,080093
X3	0,418423	0,012943	1,000000	0,014737	-0,078308
X4	0,379933	0,089421	0,014737	1,000000	-0,227991
X5	0,222500	0,080093	-0,078308	-0,227991	1,000000

Testování hypotézy o shodě vektorů středních hodnot pomocí Hotellingova testu

Proměnná	t-testy; grupováno: ID1 (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	F-poměr Rozptyly	p Rozptyly
X1	42,84483	59,76190	-7,40751	48	0,000000	29	21	7,013894	9,142783	1,699176	0,193069
X2	4,24138	5,14286	-1,89805	48	0,063712	29	21	1,661651	1,651839	1,011916	0,995884
X3	4,27586	5,76190	-3,15623	48	0,002760	29	21	1,623412	1,670472	1,058816	0,872933
X4	3,72414	4,33333	-1,73042	48	0,089980	29	21	1,130630	1,354006	1,434168	0,372786
X5	46,93103	53,61905	-3,01289	48	0,004122	29	21	7,568407	7,990471	1,114643	0,776989

Simultánní testy o složkách vektorů středních hodnot

Proměnná	t-testy; grupováno: ID1 (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	Sm.odch. návštěva ne	Sm.odch. návštěva ano	T0j =2,232*(v1-	kvantil =VF(0,95;5
X1	42,84483	59,76190	7,013894	9,142783	10,4741845	2,42704012
X2	4,24138	5,14286	1,661651	1,651839	0,68768495	2,42704012
X3	4,27586	5,76190	1,623412	1,670472	1,90157282	2,42704012
X4	3,72414	4,33333	1,130630	1,354006	0,57158454	2,42704012
X5	46,93103	53,61905	7,568407	7,990471	1,73277887	2,42704012

Vidíme, že uvažované dvě skupiny rodin se liší především v proměnné X_1, tj. v příjmu.

Provedení kanonické diskriminační analýzy pro dvě skupiny

Získání charakteristických čísel a charakteristických vektorů pro stanovení kanonických proměnných (vzhledem k tomu, že máme jen dvě skupiny, stačí pouze 1. kanonická proměnná)

Statistiky – Vícerozměrné průzkumné techniky – Diskriminační analýza – Proměnné – Grupovací: ID1, Seznam nezáv. Proměnných: X1 až X5 – OK – Detaily – Kanonická analýza – Detaily – Výpočet: Test chí kvadrát postupných kořenů. Dostaneme tabulku:

	Test chí-kvadrát po odstranění post. kořenů (dovolena.sta)					
Kořeny odstraněny	Vlastní číslo	Kan. R	Wilk. Lambda	Chi-kv.	sv	p-hodn.
0	1,615846	0,785948	0,382285	43,75224	5	0,000000

Z této tabulky nás zajímá první sloupec, v němž je uvedena hodnota největšího charakteristického čísla matice $\mathbf{BE}^{-1}$, tedy $\lambda_1 = 1{,}6158$

Výpočet standardizovaných a prostých koeficientů 1. kanonické proměnné

Návrat do Kanonické analýzy – Koeficienty pro kanonické proměnné

Prosté koeficienty

Proměnná	Prosté koeficienty (dovolena.sta) pro kanonické proměnné Kořen1
X1	0,10639
X2	0,10128
X3	0,16523
X4	-0,02915
X5	0,06050
Konstant	-9,48474
Vlastní	1,61585
KumPodíl	1,00000

Standardizované koeficienty

Proměnná	Standardiz. koeficienty (dovolena.sta) pro kanonické proměnné Kořen1
X1	0,847952
X2	0,167872
X3	0,271498
X4	-0,035819
X5	0,468679
Vlastní	1,615846
KumPodíl	1,000000

Vyjádření 1. kanonické proměnné:

$Y_1 = -9{,}4847 + 0{,}1064\ X_1 + 0{,}1013\ X_2 + 0{,}1652\ X_3 - 0{,}0292\ X_4 + 0{,}0605\ X_5$

Podle velikosti standardizovaných koeficientů lze soudit, že největší vliv na 1. kanonickou proměnnou má X_1 (příjem), podstatně menší X_5 (věk nejstaršího člena), dále X_3 (význam rodinné dovolené), X_2 (postoj k cestování) a nejmenší pak X_4 (počet členů rodiny).

Získání koeficientů korelace mezi jednotlivými proměnnými a 1. kanonickou proměnnou
Návrat do Kanonické analýzy – Faktorová struktura

	Faktorová strukturní matice (dovolena.sta) Korelační proměnné - Kanonické kořeny (vnitřní korelace)
Proměnná	Kořen1
X1	0,841108
X2	0,215519
X3	0,358384
X4	0,196486
X5	0,342108

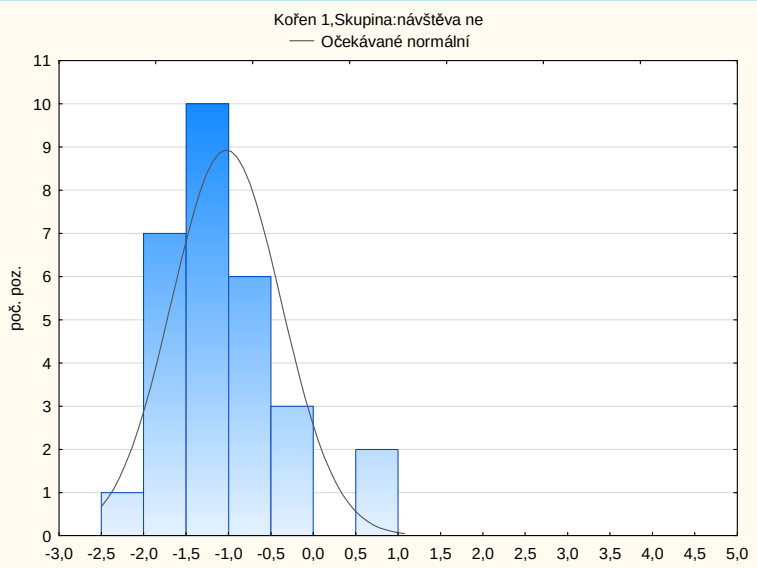
Nejvyšší korelaci pozorujeme u proměnné X_1, tedy pro 1. kanonickou proměnnou je charakteristický příjem rodiny.

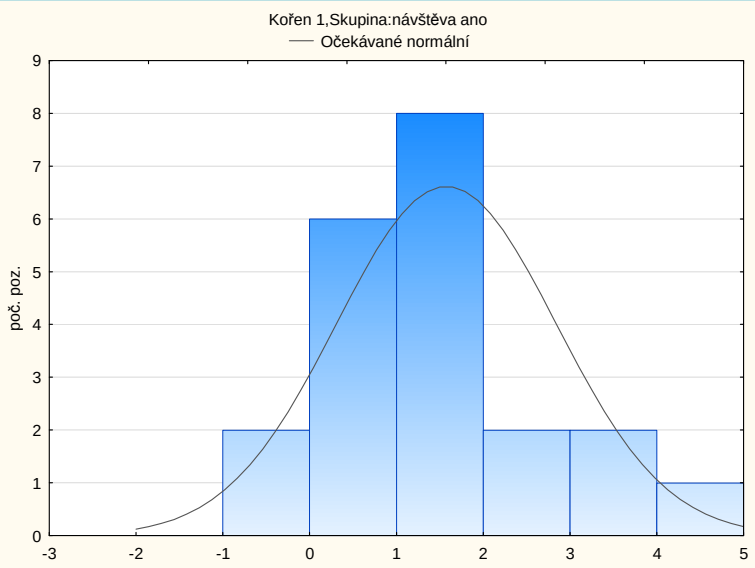
Výpočet kanonických skóre jednotlivých objektů
Návrat do Kanonické analýzy – záložka Kanonická skóre – Kanonická skóre pro každý případ

	Nestandardizovaná kanonická skóre (dovolena.sta)	
Případ	**Skup.**	Kořen1
1	návštěva ne	-1,56843
2	návštěva ne	-1,70974
3	návštěva ne	-1,46362
4	návštěva ne	-0,47001
5	návštěva ne	-1,16412
6	návštěva ne	-1,18568
7	návštěva ne	-0,74054
8	návštěva ne	-1,14695
9	návštěva ne	0,62579
10	návštěva ne	0,55324

Zobrazení jednotlivých objektů v kanonickém prostoru není v tomto případě možné, protože kanonickou diskriminační analýzou jsme 5-rozměrná pozorování zobrazili prostřednictvím 1. kanonické funkce na přímku.

Zobrazení histogramů kanonických skóre v 1. a 2. skupině
Návrat do Kanonické analýzy – záložka Kanonická skóre – Histogram kanonických skóre.
Následně upravíme měřítko na vodorovné ose.





Výpočet skupinových centroidů 1. kanonické proměnné
Návrat do Kanonické analýzy – záložka Detaily – Průměry kanonických proměnných

	Průměry kan. proměnných (dovolena.sta)
Skup.	Kořen1
návštěva ne	-1,05985
návštěva ano	1,46361

Zařazování objektů do skupin

Protože máme jenom jednu kanonickou proměnnou, můžeme určit hraniční bod:

$$C = \frac{-1{,}05985 + 1{,}46361}{2} = 0{,}202,$$

Podle něhož rozdělíme objekty do dvou skupin. Objekty, jejichž kanonické skóre je menší než C, zařadíme do 1. skupiny a ostatní objekty do 2. skupiny.

V tabulce s daty vytvoříme dvě nové proměnné skore a zarazeni. Do Dlouhého jména proměnné skore napíšeme
=0,10639*X1+0,10128*X2+0,16523*X3-0,02915*X4+0,0605*X5-9,48474
A do Dlouhého jména proměnné zarazeni napíšeme
=iif(skore > 0,202;1;0)
V proměnné skore jsou uložena kanonická skóre jednotlivých objektů a v proměnné zarazeni dostaneme zařazení objektů do skupin podle jejich kanonického skóre:

	ID1	ID2	X1	X2	X3	X4	X5	skore	zarazeni
1	návštěva ne	malá	32,1	5	4	6	58	-1,5682	0
2	návštěva ne	střední	40,0	4	4	3	42	-1,70955	0
3	návštěva ne	malá	36,2	4	3	2	55	-1,46341	0
4	návštěva ne	střední	43,2	2	5	2	57	-0,46978	0
5	návštěva ne	střední	50,4	5	2	4	37	-1,16392	0
6	návštěva ne	střední	45,2	4	4	4	42	-1,18547	0

Posouzení účinnosti diskriminace

Vytvoříme kontingenční tabulku proměnných ID1 a zarazeni, tj. klasifikační matici:

ID1	zarazeni 0	zarazeni 1	Řádk. součty
návštěva ne	27	2	29
návštěva ano	5	16	21
Vš.skup.	32	18	50

Na hlavní diagonále jsou správně zařazené případy, je 27+16=43, tj. $\frac{43}{50}100\% = 86\%$. Chybně tedy bylo zařazeno 2+5=7, tj. $\frac{7}{50}100\% = 14\%$ rodin.

Zařazení nového případu podle jeho kanonického skóre

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 kanonické diskriminace zařadit tuto rodinu do skupiny rodin, které buď navštěvují nebo nenavštěvují danou rekreační oblast.

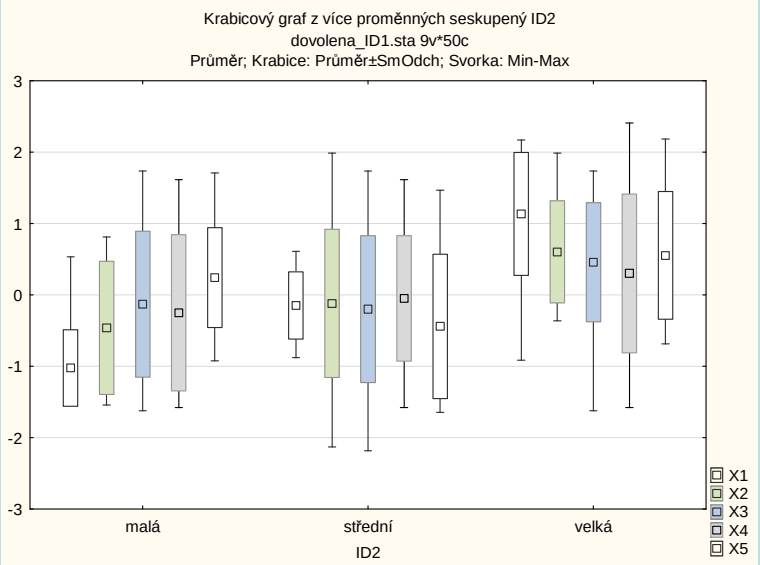
Do datové tabulky přidáme další případ a vyplníme hodnoty proměnných X1 až X5.

Přepočítáme proměnné skore a zarazeni. Proměnná skore nabývá hodnoty 0,749452, tedy tato rodina je zařazena do skupiny 1, tj. do skupiny rodin, které navštěvují danou rekreační oblast.

Řešení pro tři skupiny (třídění podle ID2):

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



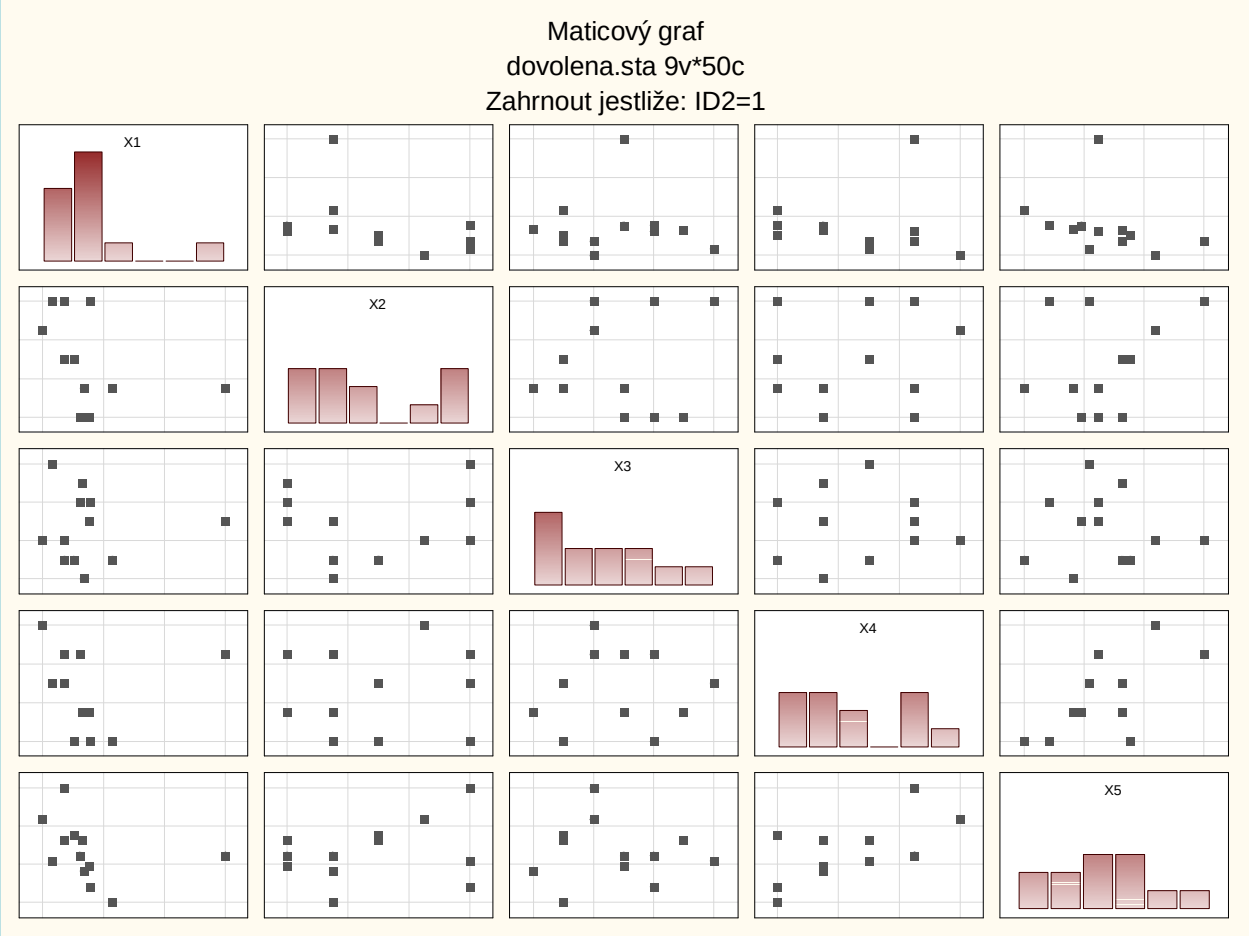
Ověření normality proměnných X_1, …, X_5 v daných třech skupinách

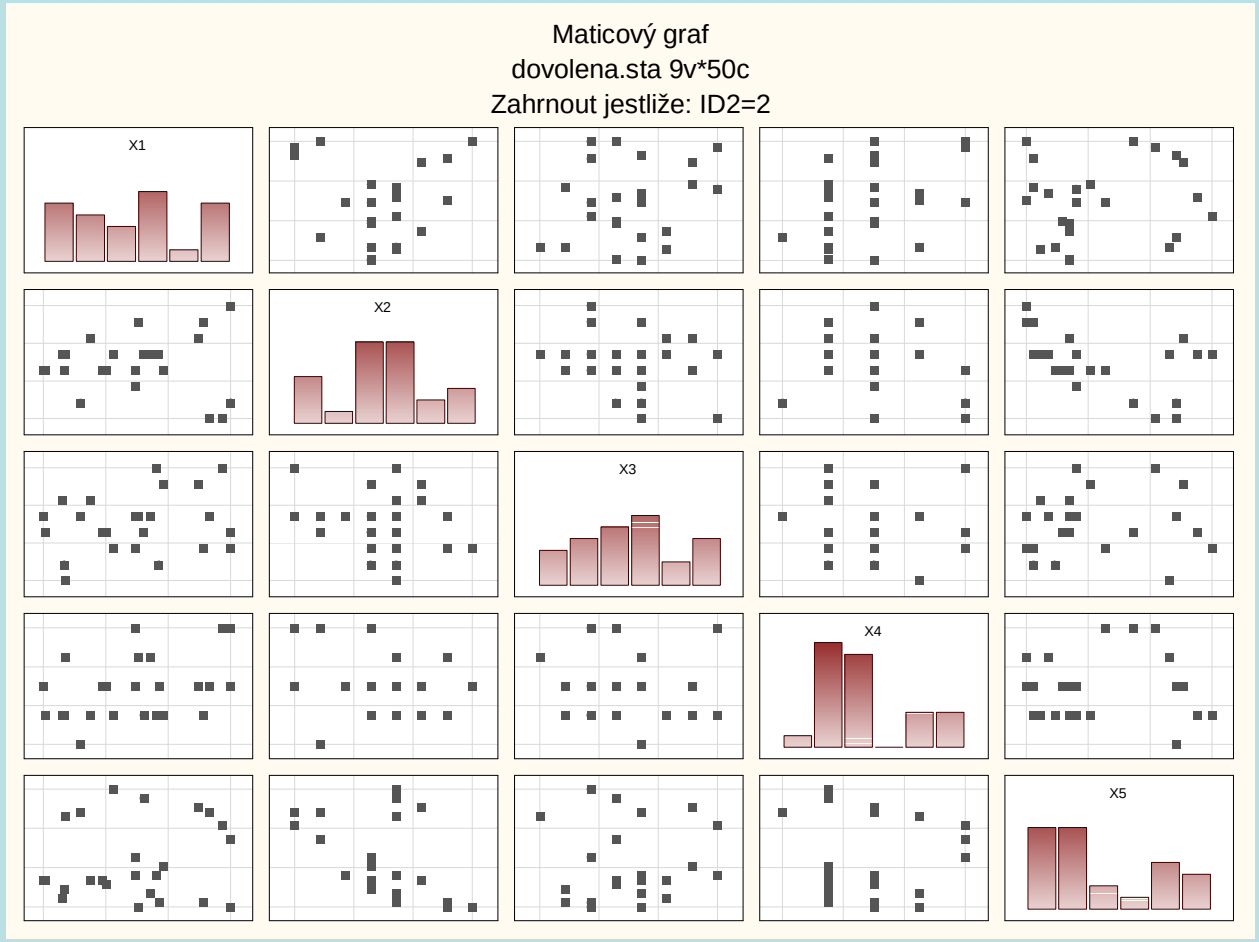
Proměnná	Souhrnné výsledky Testy normality (dovolena.sta) ID2	N	W	p
X1: roční příjem v tisících dolarů	malá	12	0,706875	0,000982
X2: postoj k cestování (škála 9 bodů)	malá	12	0,867375	0,060535
X3: význam rodinné dovolené (škála 9 bodů)	malá	12	0,955130	0,712720
X4: počet členů rodiny	malá	12	0,907871	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

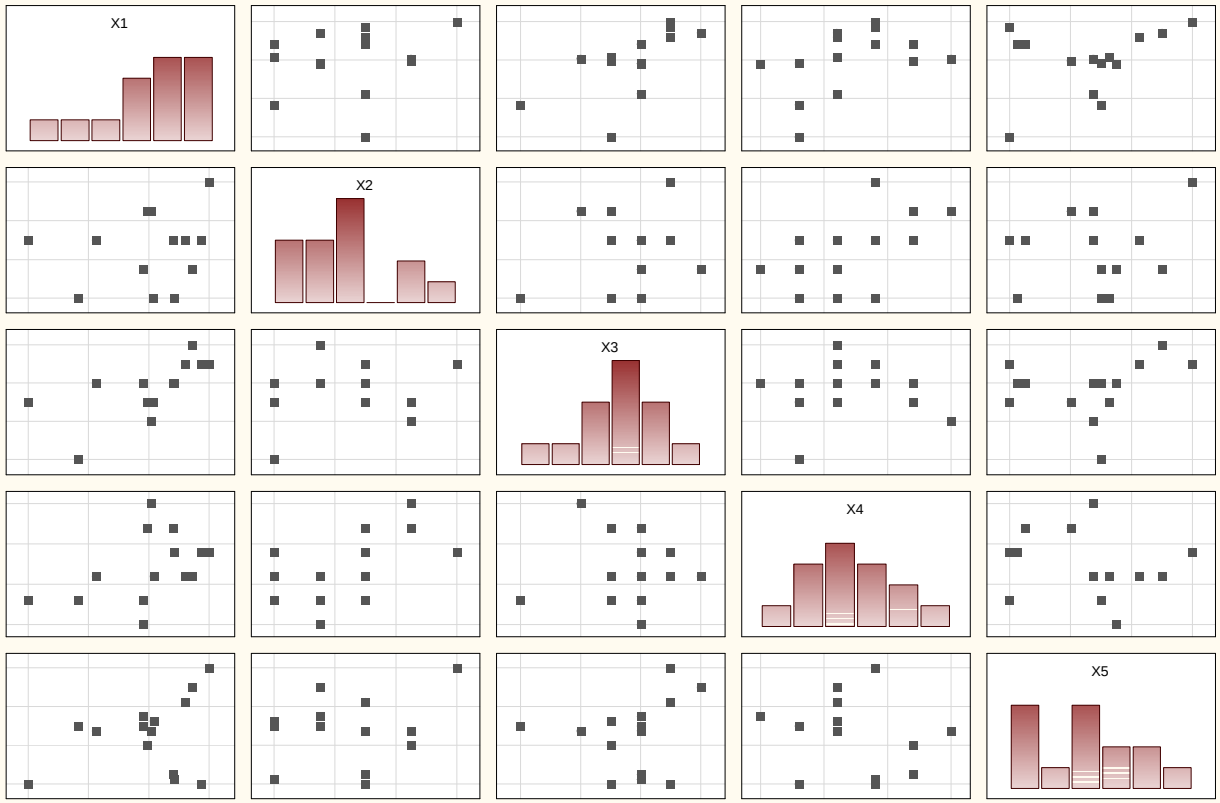
	Boxův M test (dovolena.sta) Efekt: "ID2" (Vypočteno pro všechny proměnné) Boxovo M	Chí-kv.	SV	p
Boxovo M	51,55790	42,84879	30	0,060418

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



Odhad korelační matice $\mathbf{R}_1$

Proměnná	ID2=malá Korelace (dovolena.sta)				
	X1	X2	X3	X4	X5
X1	1,000000	-0,355157	-0,016436	0,001794	-0,316554
X2	-0,355157	1,000000	0,104656	0,139401	0,298939
X3	-0,016436	0,104656	1,000000	0,133199	-0,059597
X4	0,001794	0,139401	0,133199	1,000000	0,612379
X5	-0,316554	0,298939	-0,059597	0,612379	1,000000

Odhad korelační matice $\mathbf{R}_2$

Proměnná	ID2=střední Korelace (dovolena.sta)				
	X1	X2	X3	X4	X5
X1	1,000000	-0,011874	0,229254	0,396716	0,120127
X2	-0,011874	1,000000	-0,219377	-0,205006	-0,448942
X3	0,229254	-0,219377	1,000000	-0,062614	0,079421
X4	0,396716	-0,205006	-0,062614	1,000000	0,019072
X5	0,120127	-0,448942	0,079421	0,019072	1,000000

Odhad korelační matice $\mathbf{R}_3$

Proměnná	ID2=velká Korelace (dovolena.sta)				
	X1	X2	X3	X4	X5
X1	1,000000	0,209917	0,640122	0,439641	0,322421
X2	0,209917	1,000000	0,236607	0,535452	0,175842
X3	0,640122	0,236607	1,000000	0,015888	0,216954
X4	0,439641	0,535452	0,015888	1,000000	-0,190425
X5	0,322421	0,175842	0,216954	-0,190425	1,000000

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.

Provedení kanonické diskriminační analýzy pro tři skupiny

Najdeme vlastní čísla matice $\mathbf{BE}^{-1}$:

Kořeny odstraněny	Test chí-kvadrát po odstranění post. kořenů (dovolena.sta)					
	Vlastní číslo	Kan. R	Wilk. Lambda	Chi-kv.	sv	p-hodn.
0	2,059446	0,820453	0,263219	60,06468	10	0,000000
1	0,241769	0,441245	0,805303	9,74416	4	0,044965

Dále vypočítáme prosté a standardizované koeficienty 1. a 2. kanonické proměnné:

Proměnná	Prosté koeficienty (dovolena.sta) pro kanonické proměnné	
	Kořen1	Kořen2
X1	-0,141007	-0,04449
X2	-0,220270	0,15736
X3	0,060049	0,19118
X4	0,163157	0,01824
X5	-0,015944	0,12414
Konstant	7,910380	-5,68856
Vlastní	2,059446	0,24177
KumPodíl	0,894939	1,00000

Proměnná	Standardiz. koeficienty (dovolena.sta) pro kanonické proměnné	
	Kořen1	Kořen2
X1	-1,00482	-0,317069
X2	-0,35007	0,250083
X3	0,10492	0,334027
X4	0,20433	0,022838
X5	-0,12197	0,949651
Vlastní	2,05945	0,241769
KumPodíl	0,89494	1,000000

Y1 = -0,141007*X1-0,22027*X2+0,060049*X3+0,163157*X4-0,015944*X5+7,91038
Y2 = -0,04449*X1+0,15736*X2+0,19118*X3+0,01824*X4+0,12414*X5-5,68856
Z tabulky standardizovaných koeficientů plyne, že největší vliv na zařazování do skupin má X1 a X5.

Koeficienty korelace mezi jednotlivými proměnnými a dvěma kanonickými proměnnými

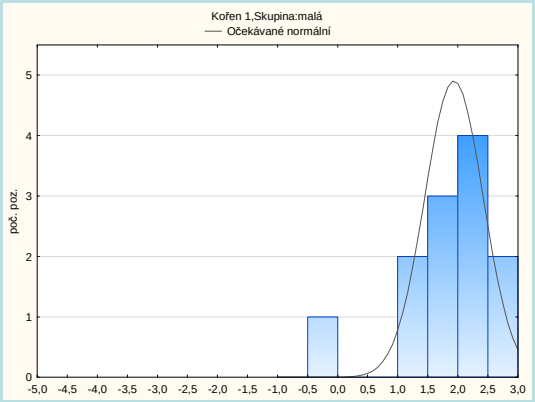
	Faktorová strukturní matice (dovolena.sta) Korelační proměnné - Kanonické kořeny (vnitřní korelace)	
Proměnná	Kořen1	Kořen2
X1	-0,918077	-0,097736
X2	-0,306332	0,065573
X3	-0,181935	0,305471
X4	-0,146634	0,009360
X5	-0,158338	0,895448

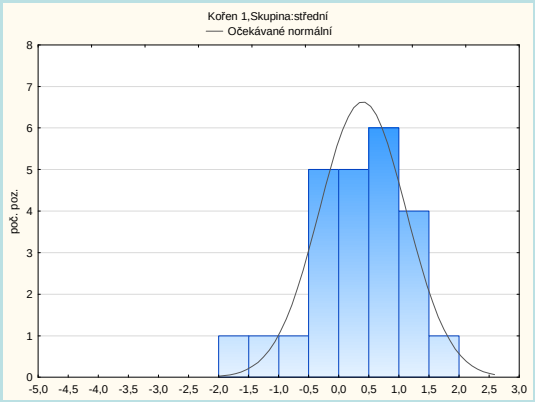
Pro 1. kanonickou proměnnou je charakteristický silný nepřímý lineární vztah s X1, zatímco pro 2. kanonickou proměnnou je charakteristický silný přímý lineární vztah s X5.

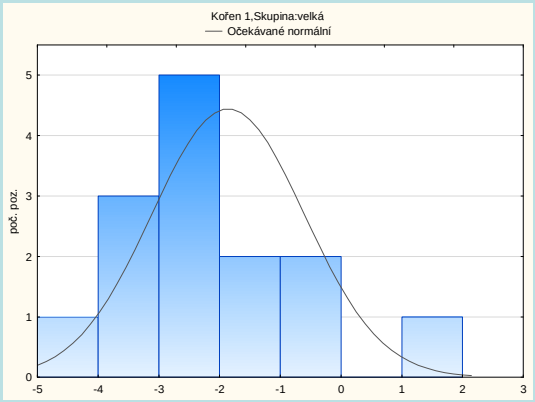
Výpočet kanonických skóre jednotlivých objektů a znázornění jejich rozmístění na ploše prvních dvou kanonických proměnných:

	Nestandardizovaná kanonická skóre (dovolena.sta)		
Případ	Skup.	Kořen1	Kořen2
1	malá	2,57711	1,74420
2	střední	1,44905	-0,80561
3	malá	1,55441	0,76788
4	střední	1,09611	0,77234
5	střední	-0,11491	-2,09582
6	střední	0,87897	-1,01875
7	střední	0,55048	-0,29098
8	malá	1,15734	0,32127
9	střední	-0,02119	0,12642
10	malá	-0,35802	-0,33442

Zobrazení histogramů kanonických skóre v 1., 2. a 3. skupině







Výpočet skupinových centroidů 1. a 2. kanonické proměnné

	Průměry kan. proměnných (dovolena.sta)	
Skup.	Kořen1	Kořen2
malá	1,74468	0,601934
střední	0,31396	-0,484385
velká	-2,03367	0,314432

Zařazování objektů do skupin není v tomto případě tak jednoduché jako v předešlé situaci, kdy jsme měli jen dvě skupiny. Zařazování se děje na základě kvadrátu Mahalanobisovy vzdálenosti kanonických skóre jednotlivých objektů od skupinových centroidů kanonických proměnných, kterou musíme vypočítat pro každou skupinu.

Návrat do Kanonická analýza – záložka Kanonická skóre – Uložit kanonická skóre – vybereme ID2 – OK. Ke vzniklé tabulce přidáme 11 nových proměnných.

V prvních šesti budou souřadnice skupinových centroidů pro 1., 2. a 3. skupinu. Nazveme je centroid11, centroid12, centroid21, centroid22, centroid31, centroid32. Do jejich Dlouhých jmen postupně napíšeme průměry kanonických proměnných, tj.
=1,74468
=0,601934
=0,31396
=-2,03367
=0,314432

Další tři proměnné nazveme d1, d2, d3 a uložíme do nich kvadráty Mahalanobisových vzdáleností kanonických skóre jednotlivých objektů od skupinových centroidů kanonických proměnných.
Do Dlouhého jména proměnné d1 napíšeme:
=(v2-v4)^2+(v3-v5)^2
Do Dlouhého jména proměnné d2 napíšeme:
=(v2-v6)^2+(v3-v7)^2
Do Dlouhého jména proměnné d3 napíšeme:
=(v2-v8)^2+(v3-v9)^2

13. proměnnou nazveme minimum a uložíme do ní nejmenší z kvadrátů Mahalanobisových vzdáleností. Do jejího Dlouhého jména napíšeme:
=min(d1;min(d2;d3))
V poslední proměnné, kterou nazveme zarazeni, bude uloženo zařazení do skupin. Vznikne překódováním proměnné minimum.
Nastavíme se kurzorem na proměnnou zarazeni – Data – Překódovat – Kategorie 1: Zahrnout, pokud v13=v10, Nová hodnota 1 – Kategorie 2: Zahrnout, pokud v13=v11, Nová hodnota 2 – Kategorie 3: Zahrnout, pokud v13=v12, Nová hodnota 3 – OK

	dovolena.sta													
	1 ID2	2 kořen1	3 kořen2	4 centroid11	5 centroid12	6 centroid21	7 centroid22	8 centroid31	9 centroid32	10 d1	11 d2	12 d3	13 minimum	14 zarazeni
1	malá	2,58	1,74	1,74468	0,601934	0,31396	-0,484385	-2,03367	0,314432	1,99772258	10,0884668	23,3035751	1,99772258	1
2	malá	1,55	0,77	1,74468	0,601934	0,31396	-0,484385	-2,03367	0,314432	0,06374111	3,10687618	13,0799182	0,06374111	1
3	malá	1,16	0,32	1,74468	0,601934	0,31396	-0,484385	-2,03367	0,314432	0,42374357	1,36036809	10,1825653	0,42374357	1
4	malá	-0,36	-0,33	1,74468	0,601934	0,31396	-0,484385	-2,03367	0,314432	5,29811437	0,47404674	3,22881309	0,47404674	2
5	malá	1,69	2,50	1,74468	0,601934	0,31396	-0,484385	-2,03367	0,314432	3,60243987	10,7923851	18,631962	3,60243987	1
6	malá	2,26	1,06	1,74468	0,601934	0,31396	-0,484385	-2,03367	0,314432	0,47726512	6,17817111	18,9885628	0,47726512	1

Posouzení účinnosti diskriminace

Vytvoříme kontingenční tabulku proměnných ID2 a zarazeni:

ID2	zarazeni2 1	zarazeni2 2	zarazeni2 3	Řádk. součty
malá	10	2	0	12
střední	3	19	2	24
velká	1	1	12	14
Vš.skup.	14	22	14	50

Na hlavní diagonále jsou správně zařazené případy: 10+19+12=41, tj. $\frac{41}{50}100\% = 82\%$. Chybně tedy bylo zařazeno 2+2+3+1+1=9, tj. $\frac{9}{50}100\% = 18\%$ rodin.