Klasifikace objektů

hledání struktury a vzájemných vazeb v objektech

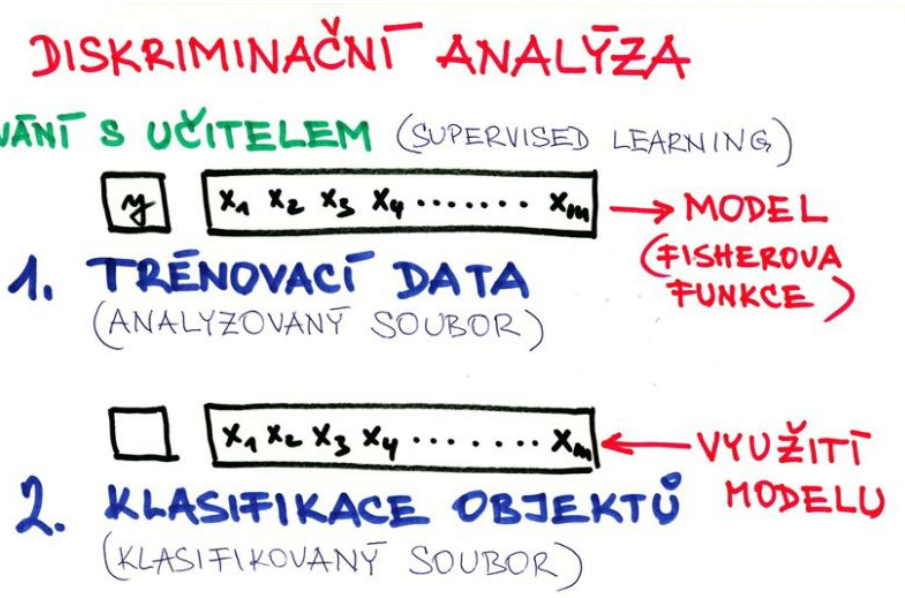
1) *Diskriminační analýza DA:* nový objekt se zařadí do již existující třídy.

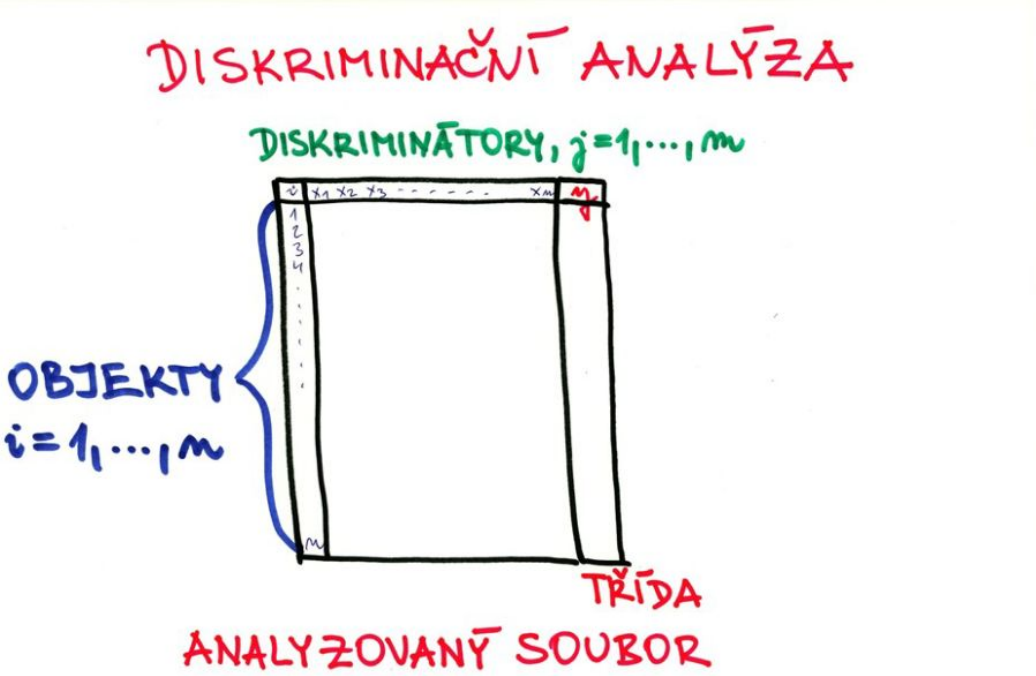
2) *Analýza shluků CLU:* neuspořádanou skupinu objektů lze uspořádat do několika vnitřně sourodých tříd či shluků

3) *Vícerozměrné škálování MDS:* hledá strukturu a vazby mezi objekty na základě jejich podobnosti.

Analyzovaný výběr jsou trénovací data, která obsahují pro každý objekt jak výstup y, tak i hodnoty všech znaků $\boldsymbol{x}$, $\{y; x_1, x_2, \ldots x_m\}$

Klasifikovaný výběr: na základě analyzovaného výběru sestavit predikční model, který umožní zařazení nových objektů do tříd.





Příklad: použití LDA pro dvě třídy $m = 2$. Vychází se ze známých matic:
$\boldsymbol{X}_1$ rozměru $n_1 \times m$ pro třídu 1, $\boldsymbol{X}_2$ rozměru $n_2 \times m$ pro třídu 2.
Jednotlivé objekty v matici $\boldsymbol{X}$ všech dat se zařadí do tříd podle výstupu $\boldsymbol{G}$.

Postup:

1) Vyčíslí se výběrové průměry $\bar{x}_1$ a $\bar{x}_2$ a společná kovarianční matice

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

2a) Nejjednodušší je předpoklad $\pi_1 = \pi_2 = 0.5$.
2b) Pokud je výběr informativní a byl pořízen jako celek a pak rozdělen do skupin, je možné použít relativních četností

$$\hat{\pi}_1 = n_1/(n_1 + n_2) \quad \text{a} \quad \hat{\pi}_2 = n_2/(n_1 + n_2)$$

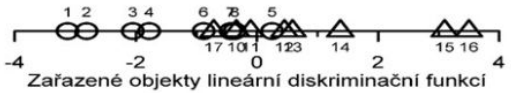
3) Za normality se určí koeficienty Fisherovy lineární diskriminační funkce z odhadů

$$\hat{\boldsymbol{a}} = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)\,\boldsymbol{S} \text{ a} \qquad \hat{b} = -0.5\,\boldsymbol{a}^T(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2) - \ln(\hat{\pi}_2/\hat{\pi}_1)$$

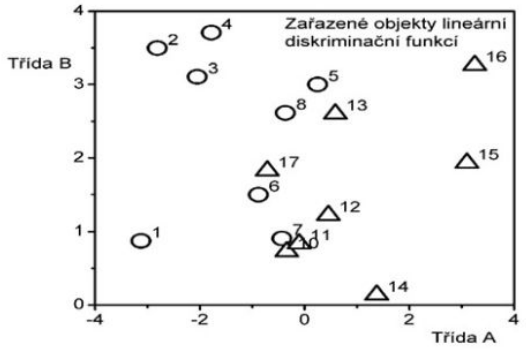
4) Při zařazování nových objektů s hodnotami znaků $\boldsymbol{x}_0$ se použije pravidlo, že do první skupiny je objekt klasifikován pokud platí $\hat{\boldsymbol{a}}^T \boldsymbol{x}_0 + \hat{b} \geq 0$.

5) V opačném případě se klasifikuje do druhé skupiny.

Ukázka 1: zařazování Fisherovou lineární diskriminační funkcí:



Ukázka 2: zařazování Fisherovou lineární diskriminační funkcí:



Příklad 4.22 ***Třídění lebek Tibeťanů lineární diskriminační funkcí***

Databáze lebek na pohřebištích v Tibetu svědčí o dvou skupinách lidí:
prvních 13 bylo nalezeno v hrobech v Sikkimu a okolí,
druhých 15 lebek na bojištích okolo Lhasy.
Předpokládejme, že máme data o 2 třídách tibetských lebek.

Data: *i* index lebky,
x1 největší délka lebky [mm],
x2 největší horizontální šířka lebky [mm],
x3 výška lebky [mm],
x4 výška horní části obličeje [mm],
x5 šířka obličeje mezi body lícních kostí [mm].

x1	*x2*	*x3*	*x4*	*x5*
190.5	152.5	145	73.5	136.5
...	...	...	...	...
...	...	...	...	...
172.5	132	125.5	63	121
167	130	125.5	69.5	119.5
1825	131	135	68.5	136

Řešení:

1) **Analyzovaný výběr:** na trénovacích datech vytvoříme třídy
První třída (13 lebek z hrobů v Sikkimu) vede ke středním hodnotám
[174.82, 139.35, 132.00, 69.82, 130.35]

a kovarianční matici $S_1 = \begin{vmatrix} 45.53 & & & & \\ 25.22 & 57.81 & & & \\ 12.39 & 11.88 & 36.09 & & \\ 22.15 & 7.52 & -0.31 & 20.94 & \\ 27.97 & 48.06 & 1.41 & 16.77 & 66.21 \end{vmatrix}$.

Druhá třída (15 lebek z bojišť v Lhase) vede ke středním hodnotám
[185.73, 138.73, 134.77, 76.47, 137.50]

a kovarianční matici $S_2 = \begin{vmatrix} 74.42 & & & & \\ -9.52 & 37.35 & & & \\ 22.74 & -11.26 & 36.32 & & \\ 17.79 & 0.70 & -10.72 & 15.30 & \\ 11.13 & 9.46 & 7.20 & 8.66 & 17.96 \end{vmatrix}$.

Koeficienty diskriminační funkce $a_1, \ldots, a_5$, jsou vyčísleny podle vztahu

$$\boldsymbol{a} = \boldsymbol{S}^{-1}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2) = [-0.09, 0.16, 0.01, -0.18, -0.18]$$

a vedou k *průměrům obou tříd*: $\bar{Z}_1 = -28.71$ a $\bar{Z}_2 = -32.21$.

Optimální prahový bod C,
dle kterého se budou nezařazené objekty třídit do první nebo druhé třídy,
se vyčíslí jako polosuma obou průměrů dle vztahu

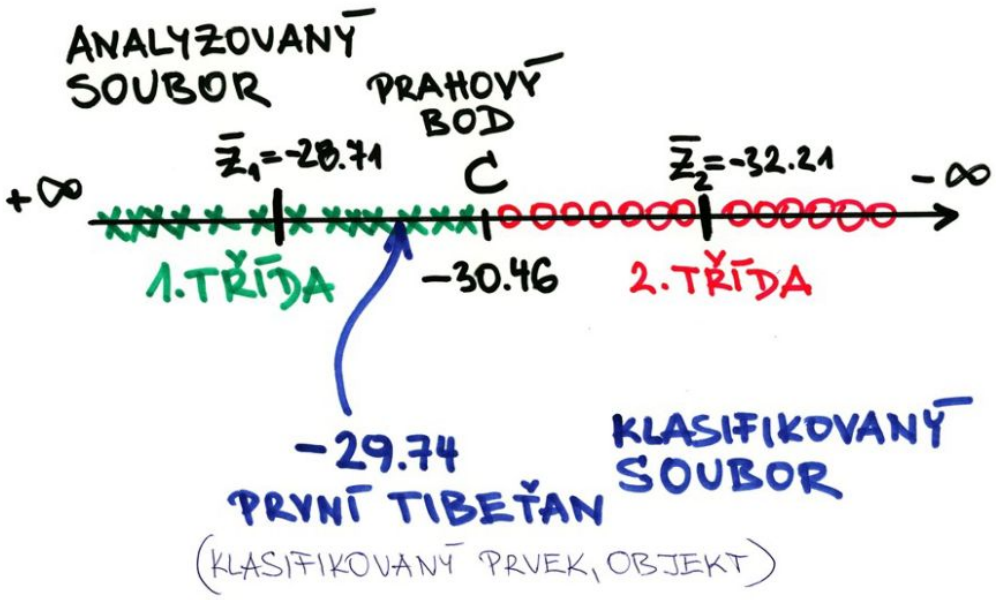
$$C = (\bar{Z}_1 + \bar{Z}_2)/2 = (-28.71 + (-32.21))/2 = -30.46.$$

Diskriminace:

2) **Klasifikovaný výběr:** na nových datech provedeme zařazení dosud nezařazených objektů
Vezmeme *data pro lebku prvního Tibeťana* a pokusíme se ji zařadit do 1. nebo do 2. třídy.
Vyčísleme proto pro ni hodnotu *lineární diskriminační funkce*

$$Z_1 = -0.09 \times 190.5 + 0.16 \times 152.5 + 0.01 \times 145.0 - 0.18 \times 73.5 - 0.18 \times 136.5 = -29.74,$$

Závěr: Protože lineární diskriminační funkce $Z_1 = -29.74$ je menší než optimální prahový bod $C = -30.46$, *patří lebka prvního Tibeťana do první třídy.*



Úprava prahového bodu

Volba prahového bodu *C* poskytuje požadovaný poměr apriorních pravděpodobností π_1 a π_2.

Optimální volba prahového bodu C je daná vzorcem $C = \frac{\bar{Z}_1 + \bar{Z}_2}{2} + \ln\frac{\pi_1}{\pi_2}$

a když bude $\pi_1 = \pi_2 = 0.5$, bude prahový bod *C* roven $C = \frac{\bar{Z}_1 + \bar{Z}_2}{2}$

Standardizované koeficienty: hodnoty koeficientů $a_1, a_2, ..., a_p$ **nejsou** přímo porovnatelné.

Relativní vliv na každou proměnnou v diskriminační funkci získáme ze *standardizovaných diskriminačních koeficientů.*

Standardizované koeficienty se vypočtou vynásobením koeficientů a_i odpovídající směrodatnou odchylkou s_i.

Volba znaků, diskriminátorů

Znaky musí zajistit přesné zařazení objektů do tříd čili diskriminaci.

Principem selekce znaků je zajištění dostatečné separability tříd a maximalizace některé zvolené míry:

-Začneme se všemi znaky. Pak se a vypouštějí se takové znaky, které provedou nedostatečnou separaci.

-V mnoha situacích může být DA užita jako exploratorní pomůcka.

-Nejprve jsou do dat zahrnuty všechny využitelné znaky. Na začátku není známo, které znaky jsou k zařazení objektů do tříd účinné.
- V diskriminační analýze místo testování změny hodnoty čtverce korelačního koeficientu R^2 přidáním nebo odebráním proměnné testujeme změnu Mahalanobisovy vzdálenosti D_M^2.

-Užívají se stejná testační kritéria jako při výběru nezávisle proměnných v lineární regresní analýze.
-*Krokový výběr znaků* kombinuje jak jejich přidávání, tak i jejich odstraňování.

-První znak, zahrnutý do modelu ve výběrovém kritériu, má největší přijatelnou hodnotu.
-*Vybírání znaků* se ukončí, když žádné další znaky nesplňují zaváděcí nebo odstraňovací kritérium.

Kritéria k vybírání znaků

Wilkovo kritérium λ: když znak poskytuje v diskriminační funkci nejmenší hodnotu Wilkova kritéria λ, je zahrnut do modelu:
- k zavedení nebo odstranění znaku je dovolen jeden krok.
- maximální počet kroků je roven dvojnásobku počtu znaků.

- ***tolerance*** je mírou stupně lineární asociace mezi znaky. Pro *i*-tý znak platí vztah $1 - R_i^2$, kde R_i^2 je čtverec vícenásobného korelačního koeficientu, když je uvažován *i*-tý znak za závisle proměnnou a když je uvažována regresní rovnice mezi tímto *i*-tým znakem a ostatními znaky.

- ***využití tolerance:*** malé hodnoty tolerance indikují, že *i*-tý znak je lineární kombinací ostatních znaků. Znaky s tolerancí menší než 0.001 nejsou do modelu zařazeny.

Testování: významnost změny Wilkova kritéria λ po zavedení znaku do modelu nebo odstranění z modelu je založena na testačním kritériu *F*.

Na začátku procesu vybírání znaků: tolerance a minimum tolerance jsou položeny rovny 1, protože v modelu zatím nejsou znaky.

N=143	Výsledky diskriminační funkční analýzy (B23) Počet prom. v modelu: 5; grupovací: B23x6-GROUP (3 skup) Wilk. lambda: ,10863 přibliž F (10,272)=55,328 p<0,0000					
	Wilk. Lambda	Parc. Lambda	F na vyj (2,136)	p-hodn.	Toler.	1-toler. R^2
B23x1-RELWT	0,117426	0,925075	5,50754	0,005012	0,745200	0,254800
B23x2-GLUFAST	0,149708	0,725594	25,71637	0,000000	0,149227	0,850773
B23x3-GLUTEST	0,199867	0,543500	57,11504	0,000000	0,143100	0,856900
B23x4-INSTEST	0,122473	0,886949	8,66731	0,000286	0,813954	0,186046
B23x5-SSPG	0,112112	0,968922	2,18109	0,116853	0,562221	0,437779

Test statistické významnosti dotyčného diskriminátoru

N=150	Výsledky diskriminační funkční analýzy (73Iris) Počet prom. v modelu: 4; grupovací: Trida (3 skup) Wilk. lambda: ,02344 přibliž F (8,288)=199,15 p<0,0000					
	Wilk. Lambda	Parc. Lambda	F na vyj (2,144)	p-hodn.	Toler.	1-toler. R^2
SEPALLENGTH	0,024976	0,938464	4,72115	0,010329	0,347993	0,652007
SEPALWIDTH	0,030580	0,766480	21,93593	0,000000	0,608859	0,391141
PETALLENGTH	0,035025	0,669206	35,59018	0,000000	0,365126	0,634874
PETALWIDTH	0,031546	0,743001	24,90433	0,000000	0,649314	0,350686

F-test významnosti každého znaku: hodnota *F* pro změnu Wilkova kritéria

λ při přidání znaku do modelu se vyčíslí dle

$$F_{\text{změny}} = \frac{n - g - p}{g - 1} \left(\frac{\frac{1 - \lambda_{p+1}}{\lambda_p}}{\frac{\lambda_{p+1}}{\lambda_p}} \right)$$

kde n je celkový počet objektů,
g udává počet tříd,
p je počet znaků,
λ_p značí Wilkovo lambda před přidáním a
λ_{p+1} je Wilkovo lambda po přidání znaku do modelu.

Testování: Do modelu je zařazen ten znak, který způsobuje nejmenší hodnotu Wilkova kritéria λ

Raovo *V* kritérium, známé jako *Lawley-Hotellingova stopa*, je

$$V = (n - g) \sum_{i=1}^{m} \sum_{j=1}^{m} w_{ij} \sum_{k=1}^{g} (\bar{x}_{ik} - \bar{x}_i)(\bar{x}_{jk} - \bar{x}_j)$$

kde p udává počet znaků v modelu,
g značí počet tříd,
n_k je velikost k-té třídy,
$\bar{x}_{ik}$ je střední hodnota i-tého znaku v k-té třídě,
$\bar{x}$ je průměr i-tého znaku pro všechny třídy kombinované dohromady a
w^*_{ij} představuje prvek inverzní kovarianční matice mezi třídami $\boldsymbol{B}^{-1}$.

Testování: čím větší jsou rozdíly mezi středními hodnotami (průměry) tříd, tím větší je hodnota Raova *V*.

Může se však stát, že znak po zařazení do modelu sníží hodnotu Raova *V*.

Mahalanobisova vzdálenost $D_{1,2}^2$ je zobecněná míra vzdálenosti mezi dvěma třídami

1 a 2 definovaná vztahem

$$D^2_{1,2} = (n - g) \sum_{i=1}^{m} \sum_{j=1}^{m} w_{ij} (\bar{x}_{i1} - \bar{x}_{i2})(\bar{x}_{j1} - \bar{x}_{j2})$$

kde m udává počet znaků v modelu,
$\bar{x}_{i1}$ je průměr i-tého znaku ve třídě 1,
w_{ij} je prvek inverzní kovarianční matice $\boldsymbol{B}^{-1}$.

Testování: kritérium všech párů tříd vyčísleno je jako první. Znak, který měl pro dvě od začátku nejtěsnější třídy, nejmenší hodnotu $D_{1,2}^2$, je zařazen do modelu.

F-testační kritérium mezi třídami testuje nulovou hypotézu H_0: *dva vektory středních hodnot tříd objektů jsou stejné.*

Kritérium je postaveno na Mahalanobisově vzdálenosti $D_{1,2}^2$.

Testační kritérium F je definováno vztahem

$$F = \frac{(n - 1 - p)\, n_1\, n_2}{p(n - 2)\,(n_1 + n_2)}\, D_{1,2}^2$$

a může být použito k výběru znaků.

Testování: V každém kroku je zařazen do modelu ten znak, který vykazuje největší hodnotu kritéria F.

Po zavedení dalších znaků do modelu jsou sledovány změny hodnoty testačního kritéria $F_{změny}$.

Proces přidávání znaků do modelu buď pokračuje, nebo je zastaven terminačním kritériem.

Vlastnosti testu:

Test je citlivý na velikost výběru, protože velké výběry jsou náchylné snadněji vykazovat statistickou významnost než výběry malé.

I když se totiž velikosti výběru zvyšují, menší klasifikační poměr bude stále významný.

Například, pro $n = 50$, $n_s = 42$ a $k = 2$ bude $q = (50 - 42 \times 2)^2/(50(2 - 1)) = 23.12$ a kritická hodnota pro $\alpha = 0.01$ je $\chi^2_{1-\alpha}(1) = 6.63$.

Závěr: Predikce jsou statisticky významnější než apriorní pravděpodobnost π_i, která uvádí správnou klasifikaci pro 50 %.

Například zvětšíme velikost výběru na 100 a klasifikační poměr zůstane 84 %, Pressovo q se zvýší na 46.24.

Statistické míry klasifikace spolehlivosti diskriminace:

1) **Pressovo q-kritérium** vyjadřuje míru porovnání počtu správných klasifikací vůči celkové velikosti výběru a počtu tříd. Vyčíslí se vztahem

$$q = \frac{(n - n_s\, k)^2}{n(k - 1)}$$

kde n je velikost výběru,
n_s udává počet objektů správně klasifikovaných a
k značí počet tříd,

Testování: vyšetřuje diskriminační sílu klasifikační matice v porovnání s modelem pravděpodobnosti: vypočtené q se porovnává s $\chi^2_{1-\alpha}(1)$ při dané α, a to když q překročí $\chi^2_{1-\alpha}(1)$, klasifikační matice se jeví statisticky lepší než klasifikační pravděpodobnost.

5. krok: Interpretace výsledků

Tři metody určují relativní důležitost znaku:

1) **Standardizované diskriminační koeficienty:** interpretace diskriminačních funkcí vyšetřuje znaménko a velikost *standardizovaných diskriminačních koeficientů* $\boldsymbol{a}^T = [a_0, a_1, \ldots, a_p]$, které představují relativní příspěvek svého znaku do Fisherovy lineární diskriminační funkce:

Diskuze:

1) Znaky *s relativně velkými koeficienty* přispívají více do diskriminační síly diskriminační funkce než znaky s menšími koeficienty.

2) Znaménko ukazuje, že znak dává buď kladný, nebo záporný příspěvek.

3) Malý koeficient indikuje buď, že odpovídající znak je nevýznamný k určování vztahu nebo je neúplným vztahem, protože je zde vysoký stupeň multikolinearity.

4) Problémem je také značná nestabilita diskriminačních koeficientů.

Proměnná	Standardiz. koeficienty (73Iris) pro kanonické proměnné	
	Kořen1	Kořen2
SEPALLENGTH	0,42695	0,012408
SEPALWIDTH	0,52124	0,735261
PETALLENGTH	-0,94726	-0,401038
PETALWIDTH	-0,57516	0,581040
Vlastní	32,19193	0,285391
KumPodíl	0,99121	1,000000

6. krok: Ověření výsledků

Konečné stadium DA se týká potvrzení diskriminačních výsledků.

DA má tendenci přeceňovat hit poměr.

Metoda dělení do skupin (cross-validation) je velmi užitečná.

Rozdělení výběru:

1) Soubor je náhodně rozdělen *na analyzovaný výběr* a *na klasifikovaný výběr*.

2) Místo dělení na *analyzovaný* a *klasifikovaný* můžeme zcela náhodně rozdělit soubor několikrát.

Testujeme potvrzení diskriminační funkce pomocí *klasifikační matice* a *hit poměru.*

Když *najdeme znaky*, které mají největší vliv na diskriminaci mezi třídami, je dalším krokem profilování charakteristiky tříd, založené na třídních průměrech.

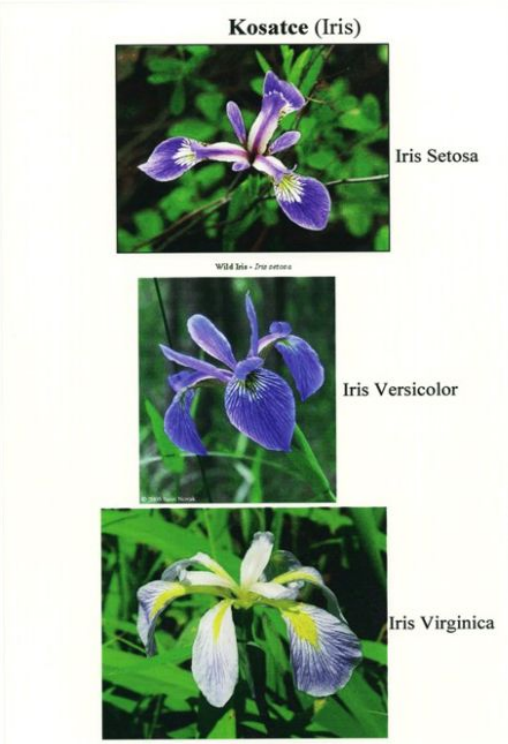
Příklad 4.24 *Botanické třídění kosatců diskriminační analýzou*

Ve Fisherově úloze o rozměrech okvětních lístků u 150 kosatců se analyzují květy tří základních tříd: (1) *Iris setosa,* (2) *Iris versicolor,* (3) *Iris virginica.* Květy kosatců jsou popsány čtyřmi znaky: délkou kališních lístků v mm anglicky *lsepal* a jejich šířkou *wsepal,* dále délkou korunních plátků v mm *lpetal* a jejich šířkou *wpetal*.

Každý objekt je popsán $p = 4$ znaky, a to *SepalLength, SepalWidth, PetalLength, PetalWidth.*

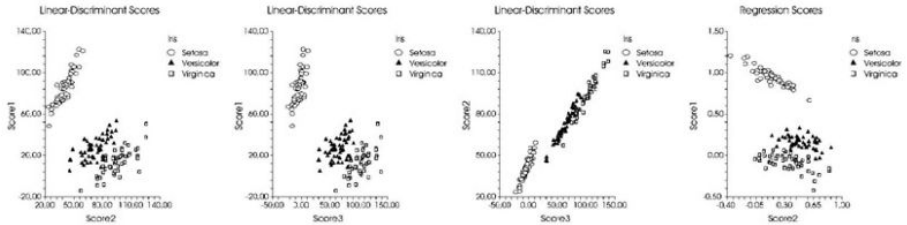
Data: Data jsou v pořadí proměnných *E418a, E418b, E418c, E418d:*

5.1	3.5	1.4	0.2	5.4	3.9	1.3	0.4	5.2	4.1	1.5	0.1	5.3	3.7	1.5	0.2
4.9	3.0	1.4	0.2	5.1	3.5	1.4	0.3	5.5	4.2	1.4	0.2	5.0	3.3	1.4	0.2
4.7	3.2	1.3	0.2	5.7	3.8	1.7	0.3	4.9	3.1	1.5	0.2	7.0	3.2	4.7	1.4
4.6	3.1	1.5	0.2	5.1	3.8	1.5	0.3	5.0	3.2	1.2	0.2	6.4	3.2	4.5	1.5
5.0	3.6	1.4	0.2	5.4	3.4	1.7	0.2	5.5	3.5	1.3	0.2	6.9	3.1	4.9	1.5
5.4	3.9	1.7	0.4	5.1	3.7	1.5	0.4	4.9	3.6	1.4	0.1	5.5	2.3	4.0	1.3
4.6	3.4	1.4	0.3	4.6	3.6	1.0	0.2	4.4	3.0	1.3	0.2	6.5	2.8	4.6	1.5
5.0	3.4	1.5	0.2	5.1	3.3	1.7	0.5	5.1	3.4	1.5	0.2	5.7	2.8	4.5	1.3
4.4	2.9	1.4	0.2	4.8	3.4	1.9	0.2	5.0	3.5	1.3	0.3	6.3	3.3	4.7	1.6
4.9	3.1	1.5	0.1	5.0	3.0	1.6	0.2	4.5	2.3	1.3	0.3	4.9	2.4	3.3	1.0
5.4	3.7	1.5	0.2	5.0	3.4	1.6	0.4	4.4	3.2	1.3	0.2	6.6	2.9	4.6	1.3
4.8	3.4	1.6	0.2	5.2	3.5	1.5	0.2	5.0	3.5	1.6	0.6	5.2	2.7	3.9	1.4
4.8	3.0	1.4	0.1	5.2	3.4	1.4	0.2	5.1	3.8	1.9	0.4	5.0	2.0	3.5	1.0
4.3	3.0	1.1	0.1	4.7	3.2	1.6	0.2	4.8	3.0	1.4	0.3	5.9	3.0	4.2	1.5
5.8	4.0	1.2	0.2	4.8	3.1	1.6	0.2	5.1	3.8	1.6	0.2	6.0	2.2	4.0	1.0
5.7	4.4	1.5	0.4	5.4	3.4	1.5	0.4	4.6	3.2	1.4	0.2	6.1	2.9	4.7	1.4



6. Klasifikační grafy.

Užijeme grafy: (a) lineárních diskriminačních skóre, (b) regresních skóre nebo (c) kanonických skóre.

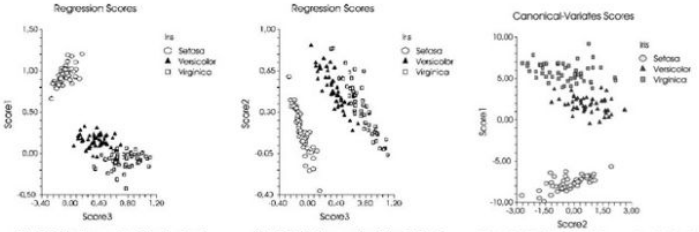


Obr. 7.9a Graf lineárního diskriminačního skóre 1 a 2 u 4 znaků 150 kosatců zdrojové matice dat *Iris* (STATISTICA).

Obr. 7.9b Graf lineárního diskriminačního skóre 1 a 3 u 4 znaků 150 kosatců zdrojové matice dat *Iris* (STATISTICA).

Obr. 7.9c Graf lineárního diskriminačního skóre 2 a 3 u 4 znaků 150 kosatců zdrojové matice dat *Iris* (STATISTICA).

Obr. 7.9d Graf regresního skóre 1 a 2 u 4 znaků 150 kosatců zdrojové matice dat *Iris* (STATISTICA).



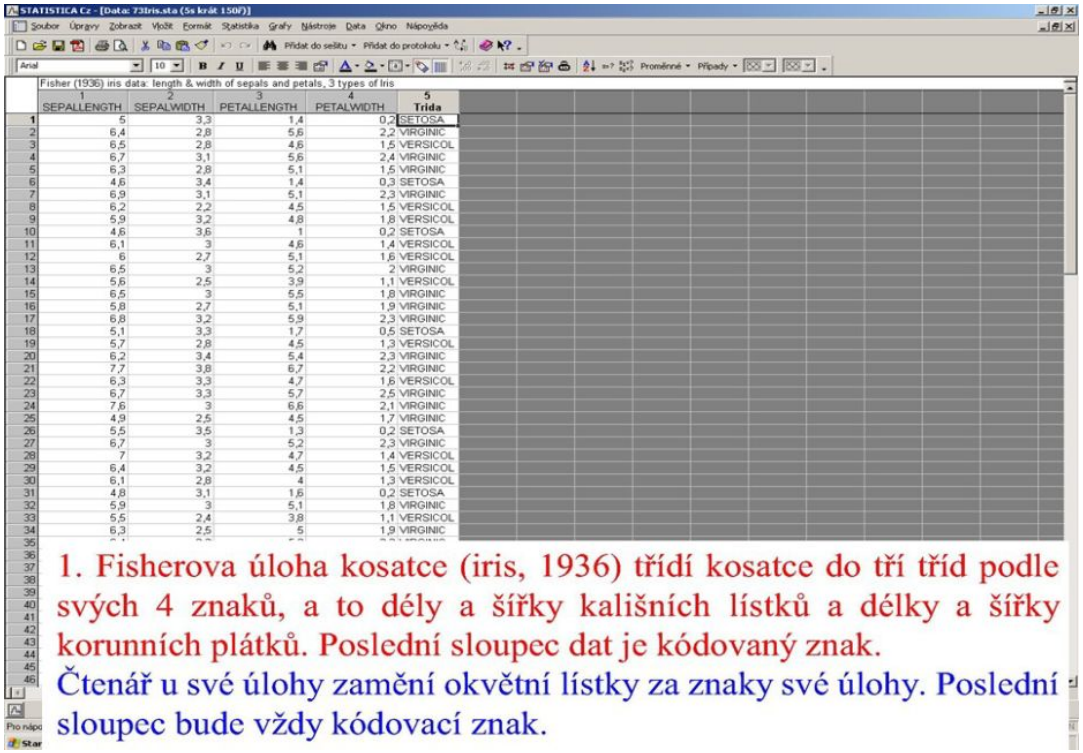
Obr. 7.9e Graf regresního skóre 1 a 3 u 4 znaků 150 kosatců zdrojové matice dat *Iris* (STATISTICA).

Obr. 7.9f Graf regresního skóre 2 a 3 u 4 znaků 150 kosatců zdrojové matice dat *Iris* (STATISTICA).

Obr. 7.9g Graf kanonických proměnných 1 a 2 u 4 znaků 150 kosatců zdrojové matice dat *Iris* (STATISTICA).

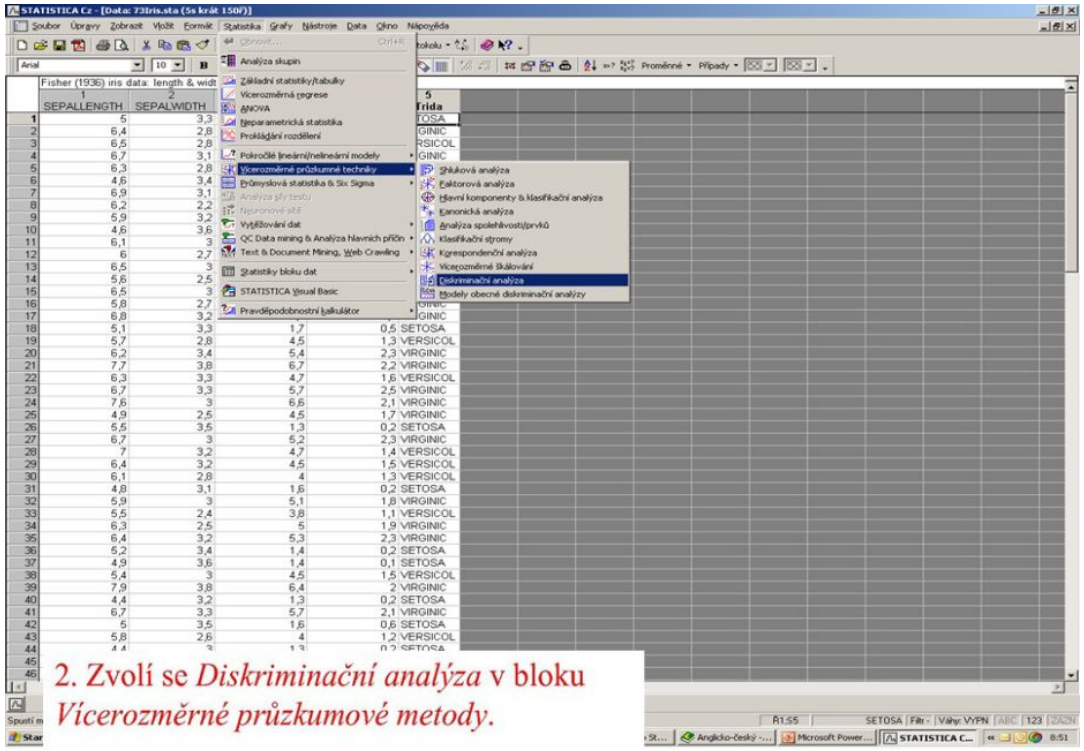
Cvičení v programu STATISTICA

1. **Načtení dat**
2. Zadání metody DA.
3. Zadání proměnných: grupovací a nezávislé.
4. Zadání kódů pro grupovací proměnné.
5. Zadání krokové analýzy
6. Definice modelu: záložky Základ, Detaily, Popisné statistiky
7. Záložka Detaily: okénko Metoda – zvol standardní (ze tří nabízených). Tolerance: 0.01
8. Záložka Popisné statistiky: volba Zobrazit popisné statistiky
9. - 38 Ze tří záložek zvol Detaily: 7 voleb postupně otevři a prohlédni.
39. STORNO a zpět do okna Popisné statistiky a zobrazit Popisné statistiky zvol záložku Všechny.
40. Zadej Celková kovariance a korelace. Korelační matice 4 znaků: Celková korelace.
44. Zvol graf korelační matice kliknutím na Graf celkové korelace.
45. Zvol Krabicový graf průměru všech 4 znaků.
46. Upravit krabicový graf.
47. 1 krát STORNO až na Definice modelu. Metoda – kroková dopředná . Záložka Detaily a pak Výsledky: pouze shrnutí.
48. Zvol Výpočet v záložce Základní nastavení. Seřadí znaky dle statistické významnosti.
50. Zpět na záložku Detaily a v ní Kanonická analýza a záložka Základní nastavení.
51. Zvol Výpočet: Test Chi kvadrát postupných kořenů. Hledání ortogonálních diskriminačních funkcí čili kanonických kořenů Chi²-testem.
52. Komentovat Chi²-test.
53. Kanonická analýza a v ní záložka Detaily a v ní záložka pro kanonické proměnné.
54. Kanonická analýza a v ní záložka Detaily a v ní Faktorová analýza a pojmenování kořenů: Kořen 1 je o délkách květů a Kořen 2 je o šířkách květů.
55. Zpět a zvol Průměry kanonických proměnných a souřadnice jejich polohy.
56. Zpět a záložka Kanonické skóre.
57.–58. Záložka Bodový graf kanonických skóre. Upravit graf.
59. Zpět a Uložit kanonická skóre za IRISTYPE a udělat Graf, Bodové grafy. Překopírovat Popisy bodů.
60. STORNO a pak záložka Klasifikace: zvol a prohlédni si záložky Klasifikační matice, Klasifikace objektů, Mahalanobisovy vzdálenosti, Aposteriorní pravděpodobnosti.



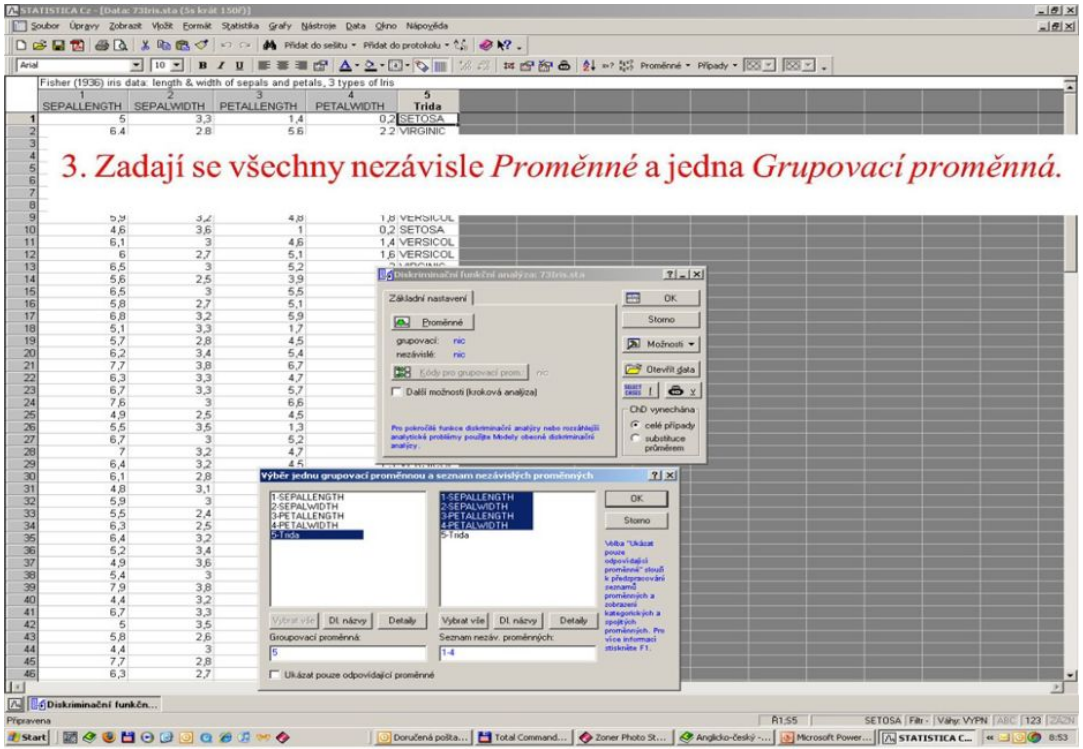
1. Fisherova úloha kosatce (iris, 1936) třídí kosatce do tří tříd podle svých 4 znaků, a to dély a šířky kališních lístků a délky a šířky korunních plátků. Poslední sloupec dat je kódovaný znak.

Čtenář u své úlohy zamění okvětní lístky za znaky své úlohy. Poslední sloupec bude vždy kódovací znak.

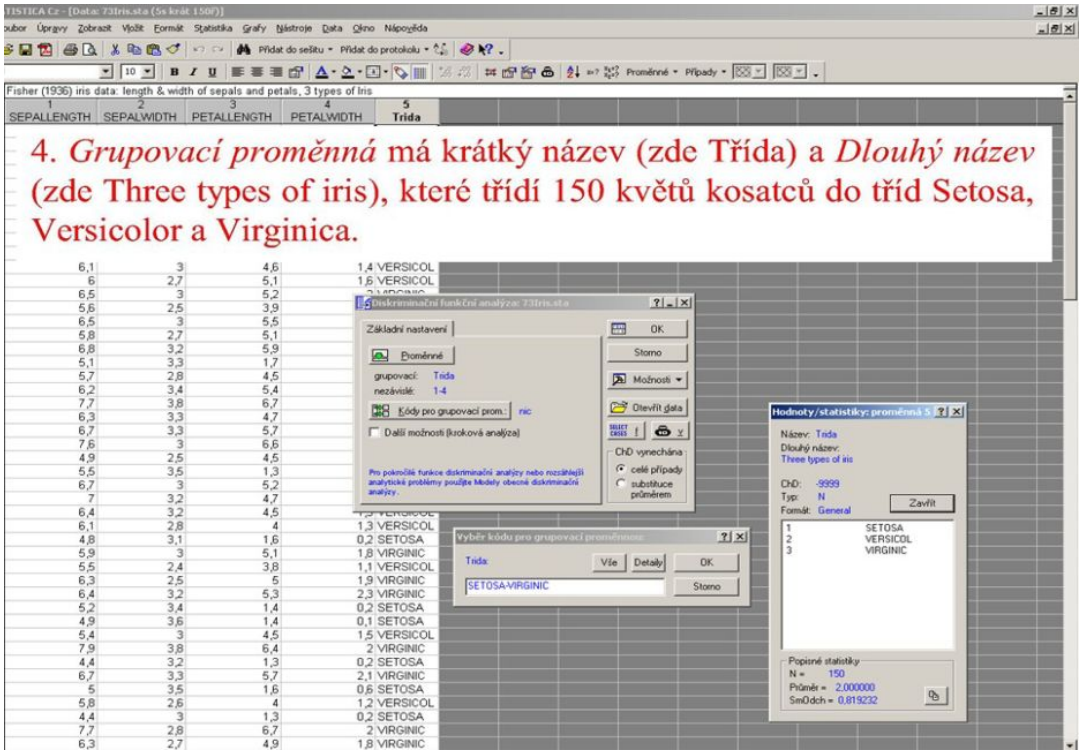


2. Zvolí se *Diskriminační analýza* v bloku *Vícerozměrné průzkumové metody.*

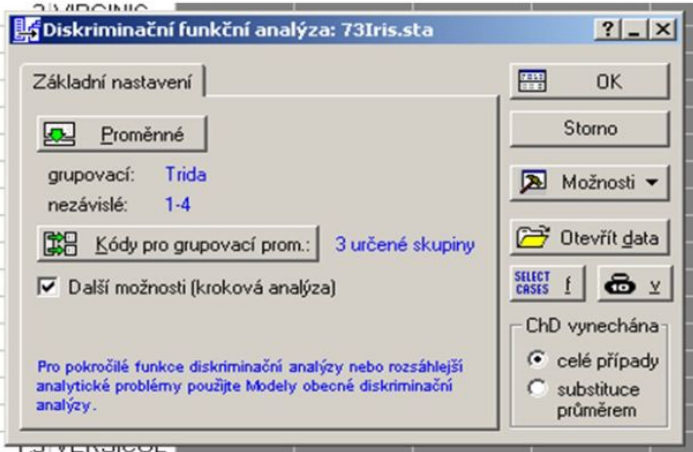
3. Zadají se všechny nezávisle *Proměnné* a jedna *Grupovací proměnná*.



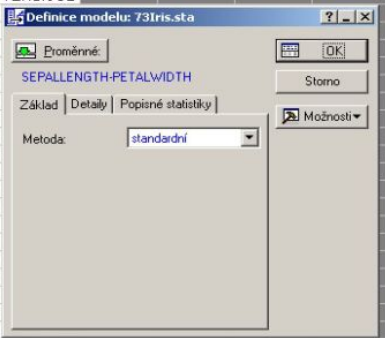
4. *Grupovací proměnná* má krátký název (zde Třída) a *Dlouhý název* (zde Three types of iris), které třídí 150 květů kosatců do tříd Setosa, Versicolor a Virginica.

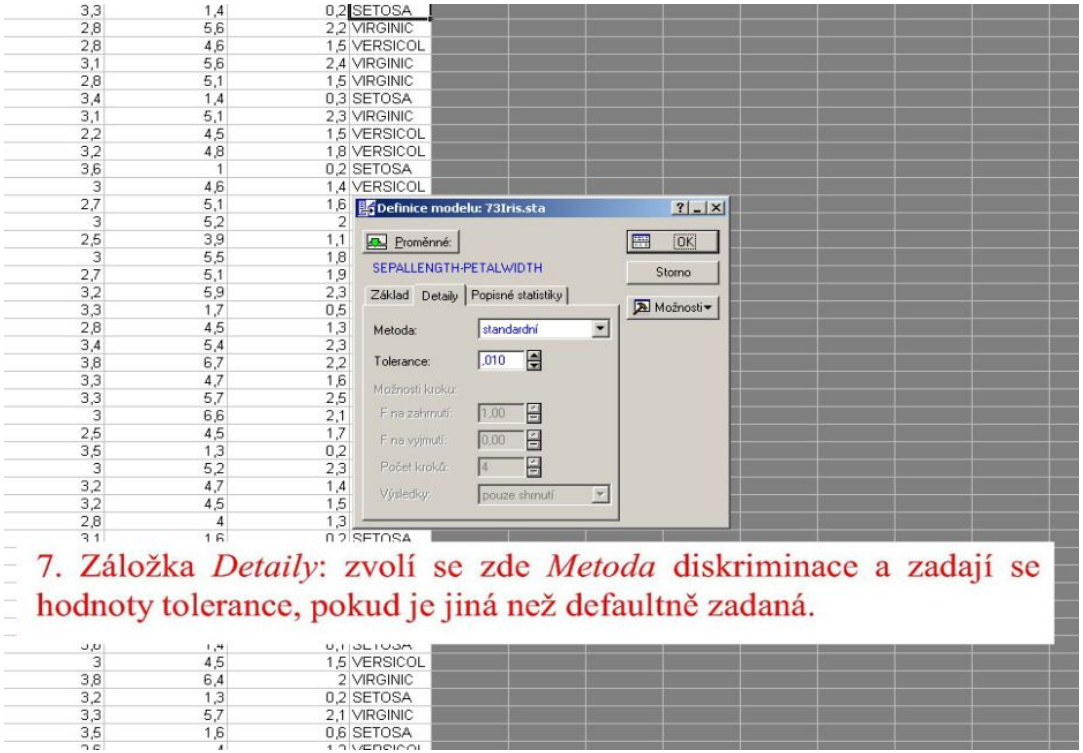


5. V okénku *Další možnosti* se zaškrtne *Kroková analýza*. Díry ve zdrojové matici vedou buď k vyškrtnutí řádku, objektu (značeno *Celé případy*) nebo k dodatečnému vyplnění díry průměrem (značeno *Substituce průměrem*).

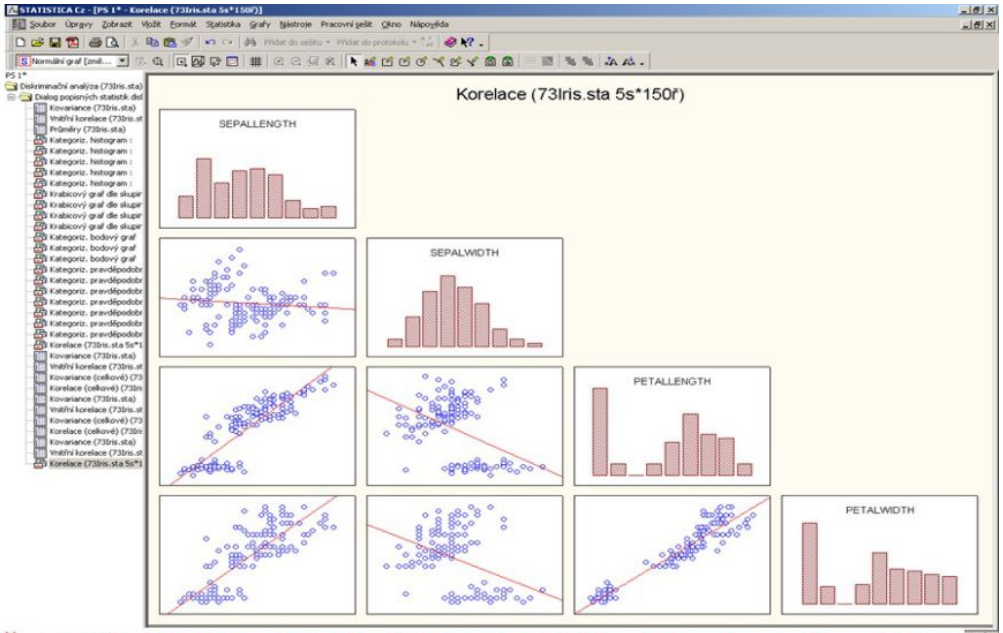


6. Definice modelu má tři záložkya to *Základ, Detaily, Popisné statistiky*, které si nyní postupně prohlédneme.





7. Záložka *Detaily*: zvolí se zde *Metoda* diskriminace a zadají se hodnoty tolerance, pokud je jiná než defaultně zadaná.



44. Červené přímky v oknech diagramu *Celkové korelace* ukazují na lineární závislost mezi dvěma znaky. Osamocené shluky bodů v každém grafu ukazují, že lze diskriminovat kosatce do několika tříd.

Výsledky diskriminační funkční analýzy (73Iris.sta)
krok 4, poč. prom. v modelu: 4; grupovací: Trida (3 skup)
Wilks. lambda: ,02344 přibliž F (8,288)=199,15 p<0,0000

N=150	Wilks. Lambda	Parc. Lambda	F na vyj (2,144)	Úroveň p	Toler.	1-toler. R^2
PETALLENGTH	0,035025	0,669206	35,59018	0,000000	0,365126	0,634874
SEPALWIDTH	0,030580	0,766480	21,93593	0,000000	0,608859	0,391141
PETALWIDTH	0,031546	0,743001	24,90433	0,000000	0,649314	0,350686
SEPALLENGTH	0,024976	0,938464	4,72115	0,010329	0,347993	0,652007

49. Program seřadí znaky a vybere přitom nejprve takový znak, který nejlépe a nejvíce přispívá k diskriminaci objektů do tříd. Jde o znak s největší hodnotou F, větší hodnotou F než bylo F zadáno na začátku, při hledání nejlepších znaků se za počáteční hodnotu totiž volí obvykle nula. Když bychom ale naopak chtěli všechny znaky z modelu odebrat, zadáme na počátku pro F vysokou hodnotu, například 9999.

Tolerance $1\text{-}R^2$ je mírou redundance (= nadbytečnosti) znaku. Vykazuje-li totiž znak toleranci třeba 0.01, pak je z 99% nadbytečný vůči dříve zadaným ostatním znakům. Když jeden nebo více znaků jsou příliš redundantní, diskriminaci nelze provést. Položíme proto mezní hodnotu tolerance rovnu 0.01.

Wilkovo λ je standardní statistika testu významnosti diskriminační síly sledovaného znaku v modelu. Nabývá hodnot od 1, značící pak žádnou diskriminační sílu znaku, do 0.0, značící perfektní diskriminační silu znaku.

Výsledky diskriminační funkční analýzy (73Iris.sta)
krok 4, poč. prom. v modelu: 4; grupovací: Trida (3 skup)
Wilks. lambda: ,02344 přibliž F (8,288)=199,15 p<0,0000

N=150	Wilks. Lambda	Parc. Lambda	F na vyj (2,144)	Úroveň p	Toler.	1-toler. R^2
PETALLENGTH	0,035025	0,669206	35,59018	0,000000	0,365126	0,634874
SEPALWIDTH	0,030580	0,766480	21,93593	0,000000	0,608859	0,391141
PETALWIDTH	0,031546	0,743001	24,90433	0,000000	0,649314	0,350686
SEPALLENGTH	0,024976	0,938464	4,72115	0,010329	0,347993	0,652007

Parciální Wilkovo λ je jednotkový příspěvek dotyčného znaku k diskriminaci objektů do tříd. Čím je jeho hodnota bližší k 0, tím větší diskriminační síla tohoto znaku. Výpočet začíná s počáteční hodnotou rovnou 0.0. Wilkovo λ konvertuje do F kritéria a odpovídající vypočtené hladiny významnosti p.

Ukazuje se zde, že *PETALLENGTH* přispívá nejvíce, *PETALWIDTH* jako druhý, pak *SEPALWIDTH* a nejméně přispívá k diskriminaci kosatců do tříd *SEPALLENGTH*. Čím menší je totiž λ, tím více přispívá znak k diskriminaci. Z toho plyne, že PETAL je hlavním znakem, který vlastně vůbec umožňuje roztřídit kosatce do tříd.

STATISTICA Cz - [PS 1* - Test chí-kvadrát po odstranění post. kořenů (73Iris.sta)]

Soubor Úpravy Zobrazit Vložit Formát Statistika Grafy Nástroje Data Pracovní sešit Okno Nápověda

PS 1*
Diskriminační analýza (73Iris.sta)
Dialog popisných statistik disl
Kovariance (73Iris.sta)
Vnitřní korelace (73Iris.st
Průměry (73Iris.sta)
Kategoriz. histogram :
Kategoriz. histogram :
Kategoriz. histogram :
Kategoriz. histogram :
Kategoriz. histogram :
Krabicový graf dle skupir
Krabicový graf dle skupir
Krabicový graf dle skupir
Krabicový graf dle skupir
Kategoriz. bodový graf
Kategoriz. bodový graf
Kategoriz. bodový graf
Kategoriz. pravděpodobr
Kategoriz. pravděpodobr
Kategoriz. pravděpodobr
Kategoriz. pravděpodobr
Kategoriz. pravděpodobr
Kategoriz. pravděpodobr
Korelace (73Iris.sta 5s*1
Kovariance (73Iris.sta)
Vnitřní korelace (73Iris.st
Kovariance (celkové) (73

Test chí-kvadrát po odstranění post. kořenů (73Iris.sta)

Kořeny odstraněny	Vlastní číslo	Kan. R	Wilks. Lambda	Chi-kv.	sv	Úroveň p
0	32,19193	0,984821	0,023439	546,1153	8	0,000000
1	0,28539	0,471197	0,777973	36,5297	3	0,000000

52. Nejprve určíme, zda obě diskriminační funkce (či kano
jsou statisticky významné. První řádek tabulky *Testu Chí-k*
statistickou významnost všech (zde 2) kanonických koř
postupném odstraňování. Druhý řádek se týká všech kořen
ale bez prvního, atd. Tabulka proto odpoví, kolik a kter
statisticky ještě významné. Zde jsou obě diskriminačn
kanonické kořeny statisticky významné. Když proto buden
u nových kosatců dvě míry okvětních lístků, PETALS
zaručíme, že je pak možné roztřídit kosatce do tříd.

STATISTICA Cz - [PS 1* - Standardiz. koeficienty (73Iris.sta)]

Soubor Úpravy Zobrazit Vložit Formát Statistika Grafy Nástroje Data Pracovní sešit Okno Nápověda

PS 1*
Diskriminační analýza (73Iris.sta)
Dialog popisných statistik disl
Kovariance (73Iris.sta)
Vnitřní korelace (73Iris.st
Průměry (73Iris.sta)
Kategoriz. histogram :
Kategoriz. histogram :
Kategoriz. histogram :
Kategoriz. histogram :
Kategoriz. histogram :
Krabicový graf dle skupir
Krabicový graf dle skupir
Krabicový graf dle skupir
Krabicový graf dle skupir
Kategoriz. bodový graf
Kategoriz. bodový graf
Kategoriz. bodový graf
Kategoriz. pravděpodobr
Kategoriz. pravděpodobr
Kategoriz. pravděpodobr
Kategoriz. pravděpodobr
Kategoriz. pravděpodobr
Kategoriz. pravděpodobr
Kategoriz. pravděpodobr
Korelace (73Iris.sta 5s*1
Kovariance (73Iris.sta)
Vnitřní korelace (73Iris.st

Standardiz. koeficienty (73Iris.sta)
pro kanonické proměnné

Proměnná	Kořen1	Kořen2
PETALLENGTH	-0,94726	-0,401038
SEPALWIDTH	0,52124	0,735261
PETALWIDTH	-0,57516	0,581040
SEPALLENGTH	0,42695	0,012408
Vlastní	32,19193	0,285391
KumPodíl	0,99121	1,000000

53. *Tabulka standardizovaných koeficientů:* Původní
spojení s původními daty poskytují skóre diskrim
Standardizované koeficienty patří ke standardizovaným
v normovaných vzájemně porovnatelných škálách. Prvn
funkce je silně vážená délkami a šířkami PETALS. Zbýv
ale také trochu přispívají. Druhá diskriminační funkce je
více znakem SEPALWIDTH a daleko však méně PE
PETALLENGTH.

STATISTICA Cz - [PS 1* - Faktorová strukturní matice (73Iris.sta)]

Soubor Úpravy Zobrazit Vložit Formát Statistika Grafy Nástroje Data Pracovní sešit Okno Nápověda

PS 1*
Diskriminační analýza (73Iris.sta)
Dialog popisných statistik disl
Kovariance (73Iris.sta)
Vnitřní korelace (73Iris.st
Průměry (73Iris.sta)
Kategoriz. histogram :
Kategoriz. histogram :
Kategoriz. histogram :
Kategoriz. histogram :
Kategoriz. histogram :
Krabicový graf dle skupir
Krabicový graf dle skupir
Krabicový graf dle skupir
Krabicový graf dle skupir
Kategoriz. bodový graf
Kategoriz. bodový graf
Kategoriz. bodový graf
Kategoriz. pravděpodobr
Kategoriz. pravděpodobr
Kategoriz. pravděpodobr
Kategoriz. pravděpodobr
Kategoriz. pravděpodobr
Kategoriz. pravděpodobr
Kategoriz. pravděpodobr
Korelace (73Iris.sta 5s*1
Kovariance (73Iris.sta)
Vnitřní korelace (73Iris.st
Kovariance (celkové) (73
Korelace (celkové) (73Iri

Faktorová strukturní matice (73Iris.sta)
Korelační proměnné - Kanonické kořeny
(vnitřní korelace)

Proměnná	Kořen1	Kořen2
PETALLENGTH	-0,706065	0,167701
SEPALWIDTH	0,119012	0,863681
PETALWIDTH	-0,633178	0,737242
SEPALLENGTH	-0,222596	0,310812

54. *Faktorová strukturní matice:* v tabulce uvede
představují korelace mezi znaky a diskriminačními fun
se proto k objasnění "významu" diskriminačních funk
výhodné pojmenovat či názvem přidělit význam těmto
funkcím podobně jako se to dělá v FA.

STATISTICA Cz - [PS 1* - Průměry kan. proměnných (73Iris.sta)]

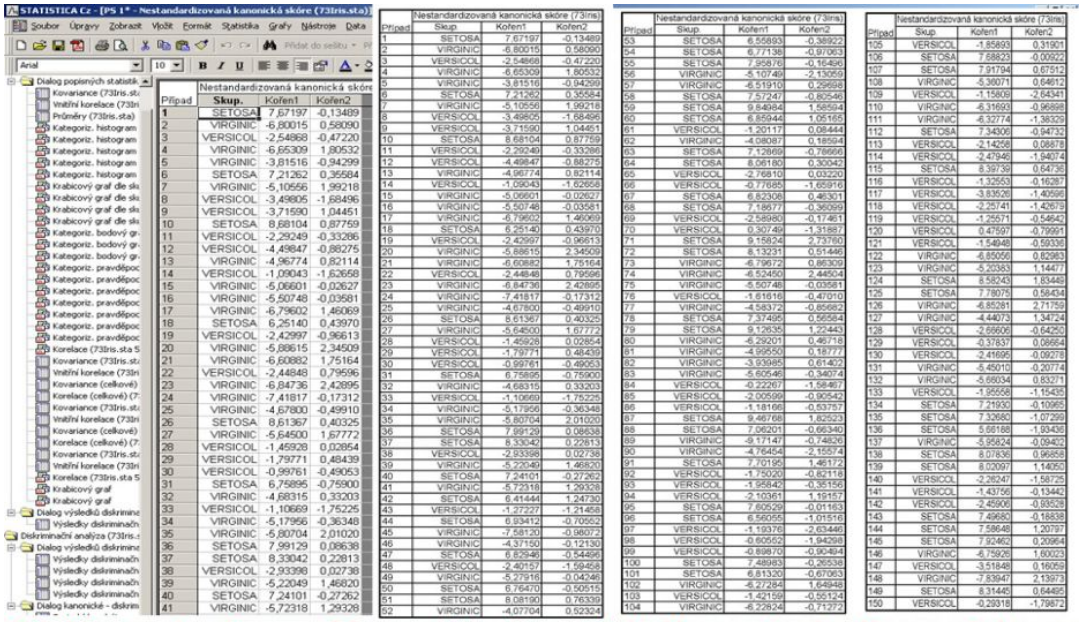
Soubor Úpravy Zobrazit Vložit Formát Statistika Grafy Nástroje Data Pracovní sešit Okno Nápověda

Diskriminační analýza (73Iris.s
Dialog popisných statistik
Kovariance (73Iris.st
Vnitřní korelace (73Iri
Průměry (73Iris.sta)
Kategoriz. histogram
Kategoriz. histogram
Kategoriz. histogram
Kategoriz. histogram
Kategoriz. histogram
Krabicový graf dle sk
Krabicový graf dle sk
Krabicový graf dle sk
Krabicový graf dle sk
Kategoriz. bodový gr
Kategoriz. bodový gr
Kategoriz. bodový gr
Kategoriz. pravděpoc
Kategoriz. pravděpoc
Kategoriz. pravděpoc
Kategoriz. pravděpoc
Kategoriz. pravděpoc
Kategoriz. pravděpoc
Kategoriz. pravděpoc
Korelace (73Iris.sta 5
Kovariance (73Iris.st
Vnitřní korelace (73Iri
Kovariance (celkové)
Korelace (celkové) (7
Kovariance (73Iris.st
Vnitřní korelace (73Iri
Kovariance (celkové)

Průměry kan. proměnných (73Iris.sta)

Skup.	Kořen1	Kořen2
SETOSA	7,60760	0,215133
VERSICOL	-1,82505	-0,727900
VIRGINIC	-5,78255	0,512767

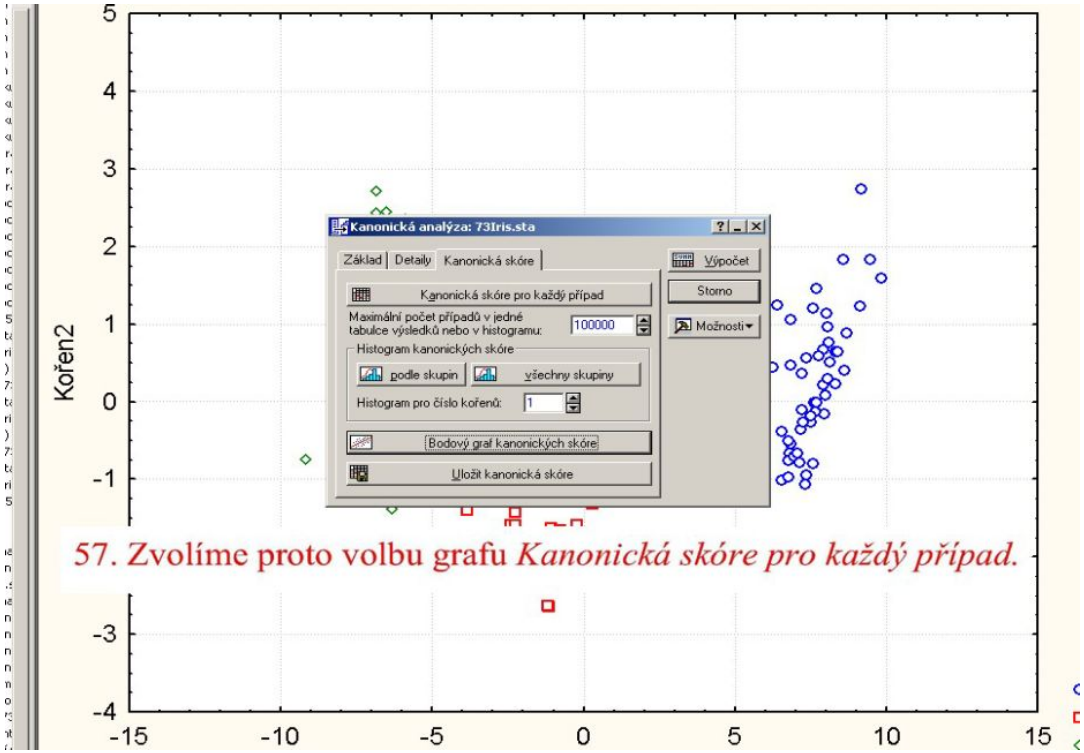
55. *Průměry kanonických proměnných:* je třeba
diskriminace kosatců v každém kanonickém kořeni. Zd
diskriminační funkce diskriminuje většinou mezi tří
ostatními kosatci. Kanonický průměr Setosy se zcela liš
dvou tříd. Druhá diskriminační funkce rozlišuje pak mezi
a ostatními kosatci, avšak tato diskriminace je mnohem
první funkce.



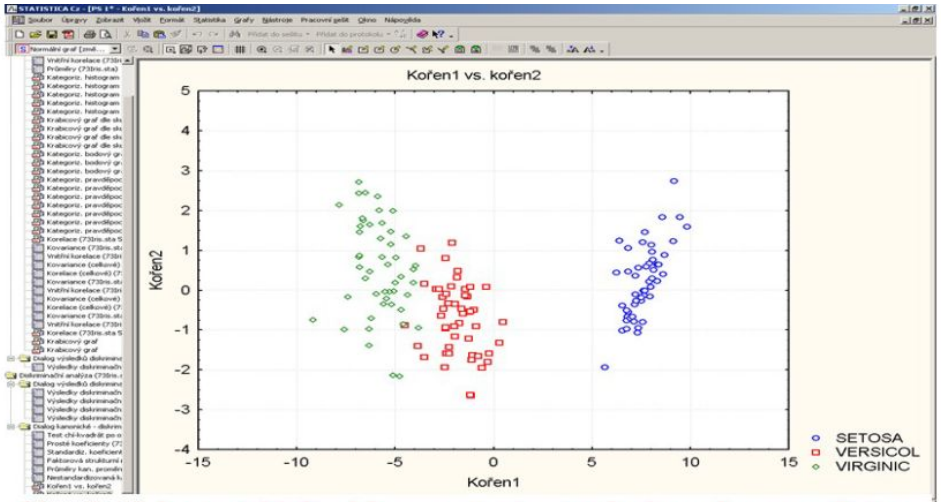
Nestandardizovaná kanonická skóre (73iris)

Případ	Skup.	Kořen1	Kořen2
1	SETOSA	7,67197	-0,13489
2	VIRGINIC	-6,80015	0,58090
3	VERSICOL	-2,54868	-0,47220
4	VIRGINIC	-6,65309	1,80532
5	VIRGINIC	-3,81516	-0,94299
6	SETOSA	7,21262	0,35584
7	VIRGINIC	-5,10556	1,99218
8	VERSICOL	-3,49805	-1,68496
9	VERSICOL	-3,71590	1,04451
10	SETOSA	8,68104	0,87759
11	VERSICOL	-2,29249	-0,33286
12	VERSICOL	-4,49847	-0,88275
13	VERSICOL	-4,96774	0,82114
14	VERSICOL	-1,09043	-1,62658
15	VERSICOL	-5,06601	-0,02627
16	VERSICOL	-5,50748	-0,03581
17	VERSICOL	-6,79602	1,46069
18	SETOSA	6,25140	0,43970
19	VERSICOL	-2,42997	-0,96613
20	VIRGINIC	-5,88615	2,34509
21	VIRGINIC	-6,60882	1,75164
22	VERSICOL	-2,44848	0,79596
23	VIRGINIC	-6,84736	2,42895
24	VIRGINIC	-7,41817	-0,17312
25	VIRGINIC	-4,67800	-0,49910
26	SETOSA	8,61367	0,40325
27	VIRGINIC	-5,64500	1,67772
28	VERSICOL	-1,45928	0,02854
29	VERSICOL	-1,79771	0,48439
30	VERSICOL	-0,99761	-0,49053
31	SETOSA	6,75895	-0,75900
32	VIRGINIC	-4,68315	0,33203
33	VERSICOL	-1,10669	-1,75225
34	VIRGINIC	-5,17956	-0,36348
35	VIRGINIC	-5,80704	2,01020
36	SETOSA	7,99129	0,08638
37	SETOSA	8,33042	0,22813
38	VERSICOL	-2,93398	0,02738
39	VIRGINIC	-5,22049	1,46820
40	SETOSA	7,24101	-0,27262
41	VIRGINIC	-5,72318	1,29328
42	SETOSA	6,41444	1,24730
43	VERSICOL	-1,27227	-1,21456
44	SETOSA	6,93412	-0,70552
45	VIRGINIC	-7,58120	-0,98072
46	VIRGINIC	-4,37150	-0,12130
47	SETOSA	6,82348	-0,54486
48	VERSICOL	-2,40157	-1,56456
49	VIRGINIC	-5,27916	-0,04246
50	SETOSA	6,76470	-0,50515
51	SETOSA	8,08190	0,76339
52	VIRGINIC	-4,07704	0,52324
53	SETOSA	6,55893	-0,38922
54	SETOSA	6,77136	-0,97063
55	SETOSA	7,95876	-0,16496
56	VIRGINIC	-5,10749	-2,13059
57	VIRGINIC	-6,51910	0,29698
58	SETOSA	7,57247	-0,80546
59	SETOSA	9,84984	1,58594
60	SETOSA	6,85944	1,05165
61	VERSICOL	-1,20117	0,08444
62	VIRGINIC	-4,08087	0,18594
63	SETOSA	7,12669	-0,78666
64	SETOSA	8,06180	0,30042
65	VERSICOL	-2,76810	0,03220
66	VERSICOL	-0,77685	-1,65916
67	SETOSA	6,82308	0,46301
68	SETOSA	7,18677	-0,36099
69	VERSICOL	-2,56980	-0,17461
70	VERSICOL	0,30749	-1,31897
71	SETOSA	9,15824	2,73760
72	SETOSA	8,13231	0,51446
73	VIRGINIC	-6,79672	0,86309
74	VIRGINIC	-6,52459	2,44502
75	VIRGINIC	-5,50748	-0,03581
76	VERSICOL	-1,61616	-0,47010
77	VIRGINIC	-4,58372	-0,85682
78	SETOSA	7,37495	0,56584
79	SETOSA	9,12636	1,72443
80	VIRGINIC	-6,29201	0,46716
81	VIRGINIC	-4,99550	0,18777
82	VIRGINIC	-3,93985	0,61402
83	VIRGINIC	-5,60546	-0,34074
84	VERSICOL	-0,22267	-1,58467
85	VERSICOL	-2,00599	-0,90542
86	VERSICOL	-1,18166	-0,53757
87	SETOSA	9,46766	1,82523
88	SETOSA	7,06201	-0,66340
89	VIRGINIC	-9,17145	-1,24563
90	VIRGINIC	-4,76454	-2,15574
91	SETOSA	7,70195	1,46172
92	VERSICOL	-1,75020	-0,82116
93	VERSICOL	-1,95842	-0,35156
94	VERSICOL	-2,10361	-1,19157
95	SETOSA	7,60529	-0,01163
96	SETOSA	6,56055	-1,01516
97	VERSICOL	-1,19376	-2,63446
98	VERSICOL	-0,60552	-1,94298
99	VERSICOL	-0,89867	-0,98464
100	SETOSA	7,48983	-0,26538
101	SETOSA	6,81320	-0,67063
102	VIRGINIC	-6,27284	1,64948
103	VERSICOL	-1,42159	-0,55124
104	VIRGINIC	-6,22924	-0,71272
105	VERSICOL	-1,85893	0,31901
106	SETOSA	7,58823	-0,00922
107	SETOSA	7,91794	0,67512
108	VIRGINIC	-5,36071	0,64612
109	VERSICOL	-1,15809	-2,64341
110	VIRGINIC	-6,31693	-0,96898
111	VIRGINIC	-6,32774	-1,38329
112	SETOSA	7,34306	-0,94732
113	VERSICOL	-2,14256	0,08878
114	VERSICOL	-2,47946	-1,94074
115	SETOSA	8,39739	0,64736
116	VERSICOL	-1,32553	-0,16287
117	VERSICOL	-3,83626	-1,40596
118	VERSICOL	-2,25741	-1,42679
119	VERSICOL	-1,26571	-0,54643
120	VERSICOL	0,47597	-0,79991
121	VERSICOL	-1,54948	-0,59336
122	VIRGINIC	-6,85056	0,82983
123	VIRGINIC	-5,20363	1,14477
124	SETOSA	8,58243	1,83449
125	SETOSA	7,78075	0,58434
126	VIRGINIC	-6,85281	2,71759
127	VIRGINIC	-4,44073	1,34724
128	VERSICOL	-2,66606	-0,64265
129	VERSICOL	-0,37837	0,08664
130	VERSICOL	-2,41695	-0,09278
131	VIRGINIC	-5,45010	-0,20774
132	VIRGINIC	-5,66034	0,83271
133	VERSICOL	-1,95558	-1,15435
134	SETOSA	7,21930	-0,10965
135	SETOSA	7,32680	-1,07299
136	SETOSA	5,66166	-1,93436
137	VIRGINIC	-5,95824	-0,09450
138	SETOSA	8,07836	0,96858
139	SETOSA	8,02097	1,14050
140	VERSICOL	-2,26247	-1,58725
141	VERSICOL	-1,43756	-0,13442
142	VERSICOL	-2,45906	-0,93528
143	SETOSA	7,49680	-0,18838
144	SETOSA	7,58648	1,20797
145	SETOSA	7,92462	0,20964
146	VIRGINIC	-6,17926	1,60020
147	VERSICOL	-3,51848	0,16059
148	VIRGINIC	-7,83947	2,13973
149	SETOSA	8,31445	0,64495
150	VERSICOL	-0,29318	-1,79872

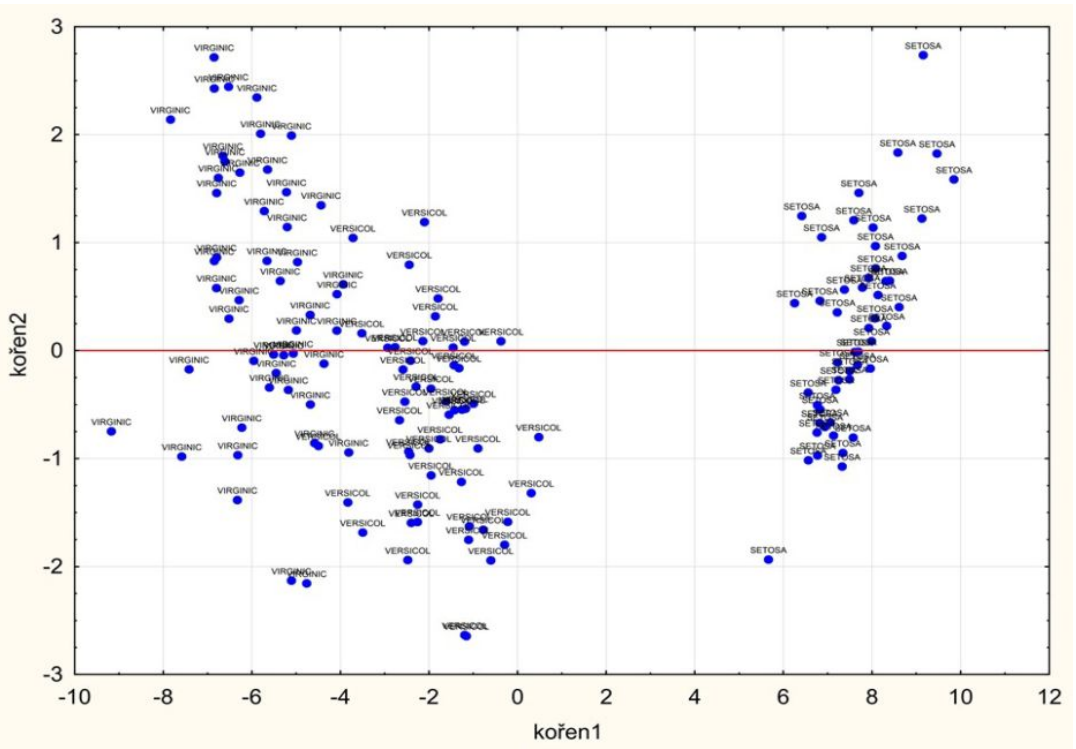
56. *Nestandardizované kanonické skóre:* Tabulka nestandardizovaných kanonických kořenů ukazuje na diskriminaci všech 150 kosatců. Samozřejmě by bylo zde názornější provést grafické zobrazení.

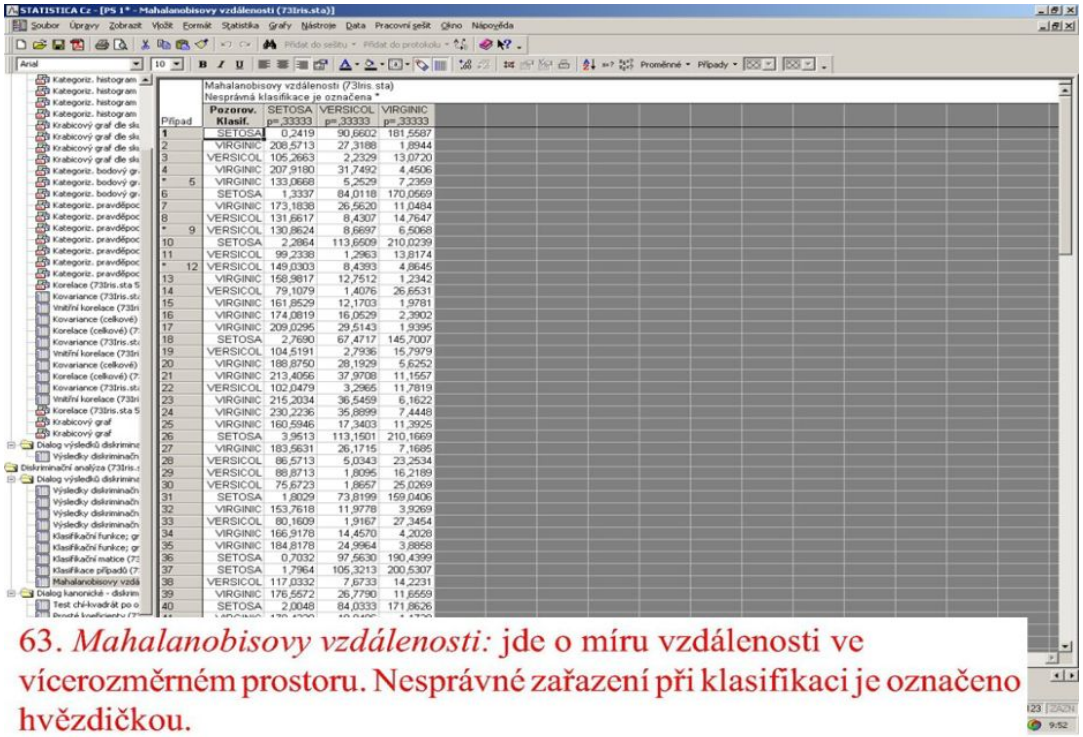


57. Zvolíme proto volbu grafu *Kanonická skóre pro každý případ.*

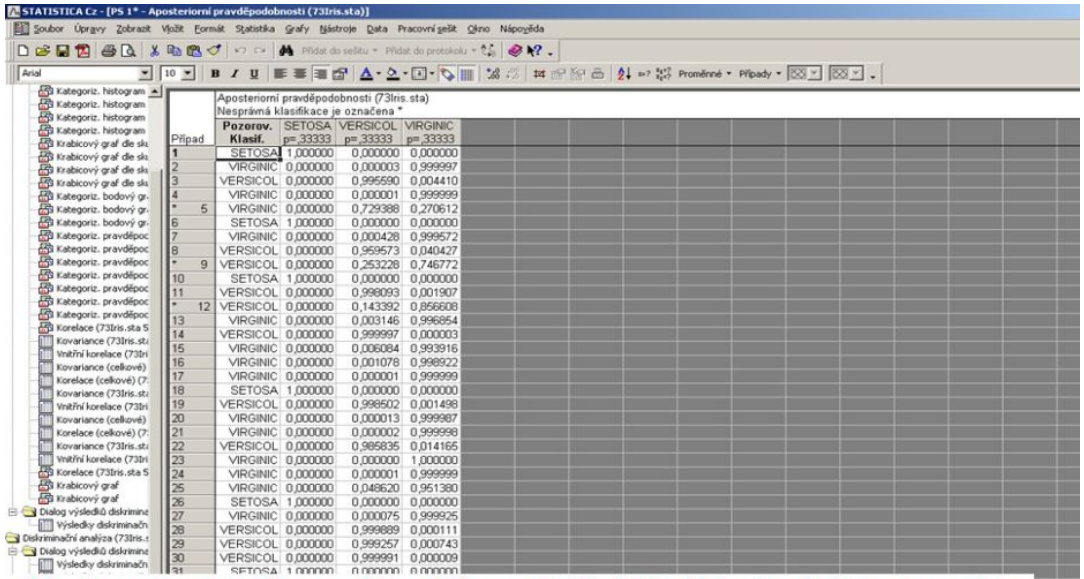


58. *Graf kanonických skóre:* v tomto grafu jsou kosatce Setosa zobrazeny modře daleko vpravo od ostatních. Kořen 1 (čili první diskriminační funkce) silně diskriminuje Setosu od ostatních kosatců. Kořen 2 (druhá diskriminační funkce) poskytuje diskriminaci mezi Versicolor se zápornými hodnotami Kořene 2 a ostatními kosatci, převážně kladného Kořene 2. Diskriminace však není tak zřetelná jako je tomu u prvního Kořene 1.

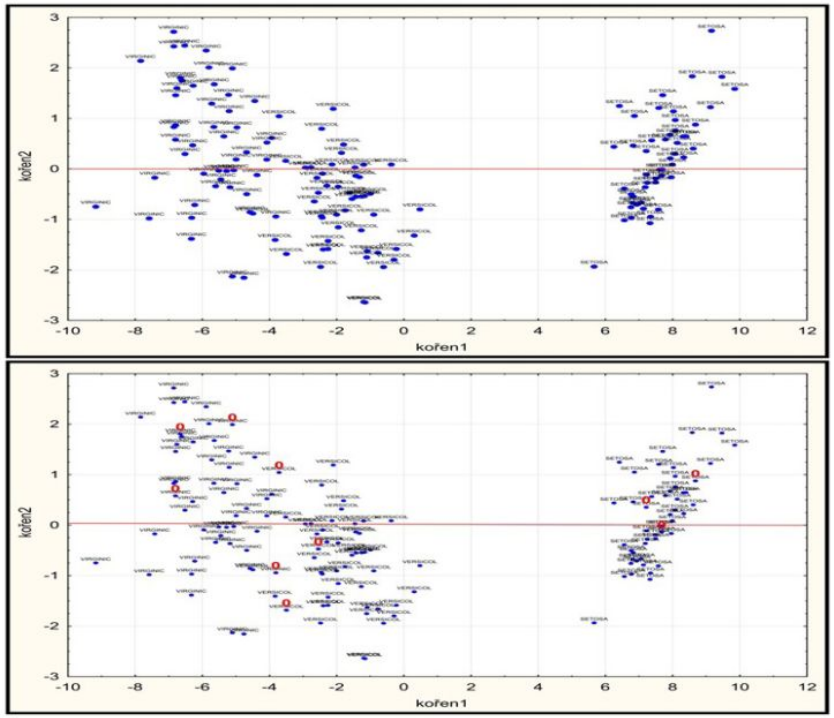




63. *Mahalanobisovy vzdálenosti:* jde o míru vzdálenosti ve vícerozměrném prostoru. Nesprávné zařazení při klasifikaci je označeno hvězdičkou.



64. *A posteriorní pravděpodobnosti:* Každá třída je dokumentována svou vypočtenou čili *a posteriorní* pravděpodobností. Největší hodnota pravděpodobnosti ve sloupečku ukazuje, že kosatec náleží do třídy dotyčného sloupce.

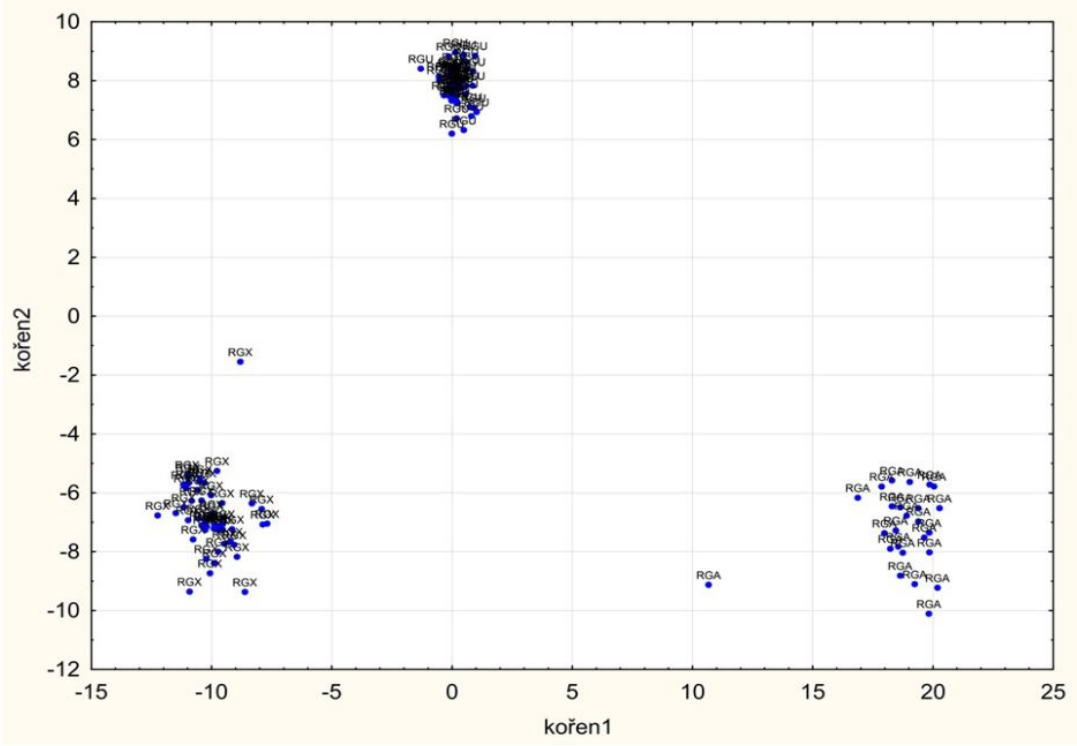


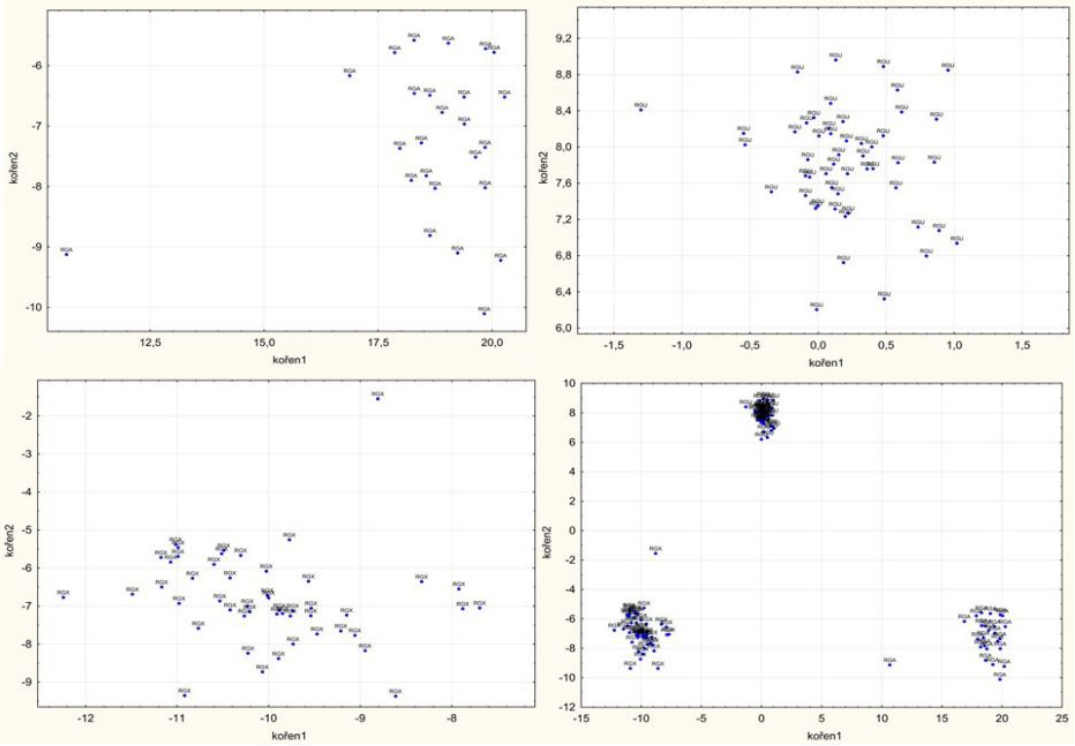
Úloha C32 ***Klasifikace a struktura znaků oxidu titaničitého dle chemických a fyzikálních vlastností*** **(DA)**

Výsledky analýz tří skupin bílého pigmentu oxidu titaničitého. Proveďte klasifikaci diskriminační analýzou (DA).

○ ***Data:*** Popis znaků ve sloupcích: Data **ID**, Zařazení do tříd **Klas**, tj. typy RGA, RG*X*, RGU tvoří tréninková data. Diskriminátory:
Ti obsah Ti (%),
TL obsah těkavých látek (%),
SpO je spotřeba oleje (g/100 g pigmentu),
Barv značí barvivost,
Podt značí podtón,
Si značí obsah Si (%),
Al značí obsah Al (%).

ID	Ti	TL	SpO	Barv	Podt	Si	Al	Klas
1	93.78	0.5	24.2	1870	10	2.174	3.511	RGA
.....	.....	.....	.....	.....	.....	.....	.....	.....
144	98.21	0.2	19	1860	13	0.354	1.014	RGX



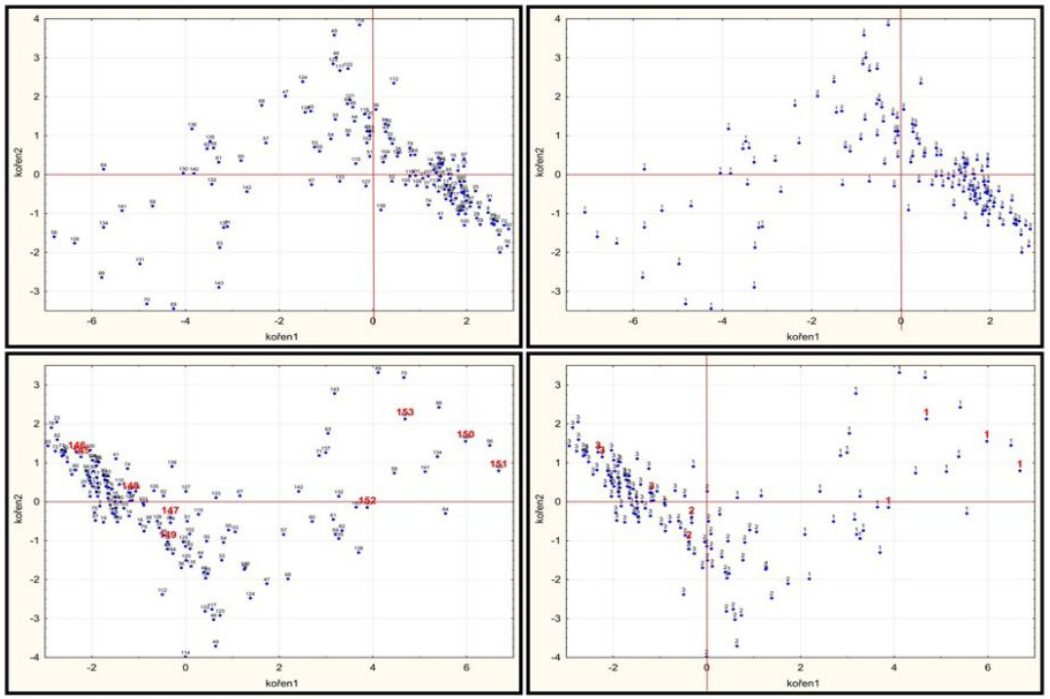


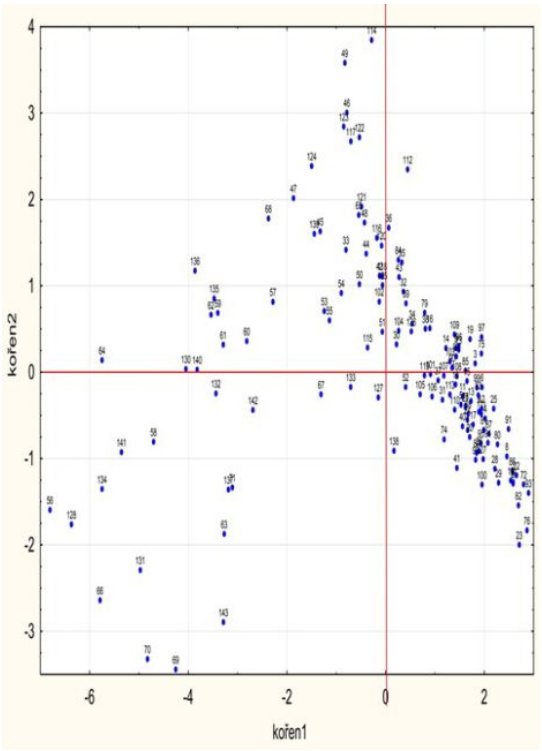
Úloha B23 *Porovnání analýzy krve diabetiků a zdravých jedinců*

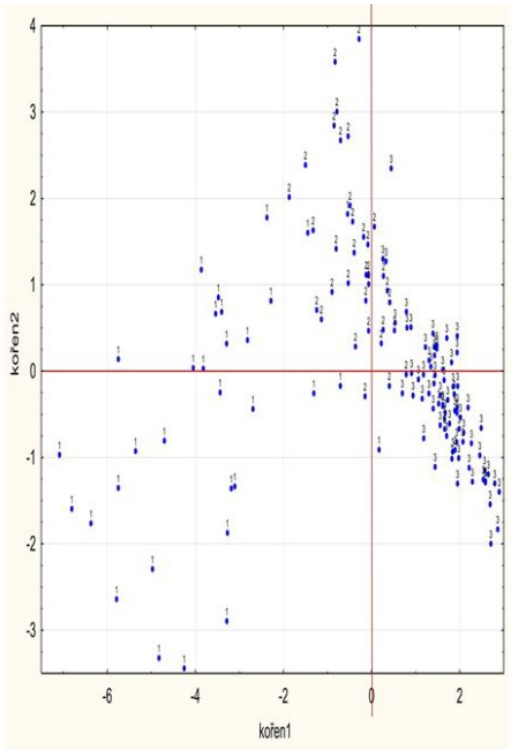
Jsou porovnány výsledky rozboru krve zdravých jedinců a dvou skupin diabetiků. Postavte soubor diskriminačních funkcí, které mohou předpovědět třídu a který rozlišuje mezi třídami podle hodnot ostatních kvantitativních proměnných.

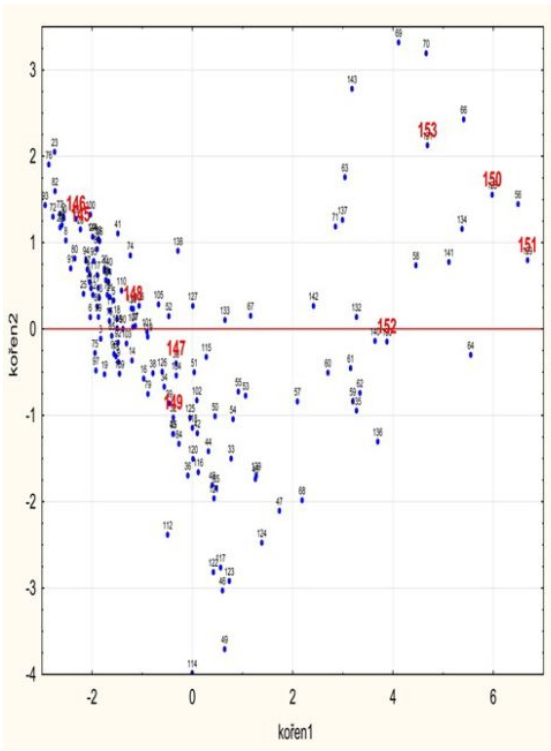
◯ ***Data:*** Výběr DIABETES se týká vyšetření 144 jedinců pro 6 znaků:
i (ID) identifikační číslo pacienta,
*x*1 (RELWT) značí relativní hmotnost pacienta,
*x*2 (GLUFAST) značí půst plasmové glukózy,
*x*3 (GLUTEST) značí prověrku plasmové glukózy,
*x*4 (INSTEST) značí plasmový inzulin v průběhu prověrky,
*x*5 (SSPG) značí ustálený rovnovážný stav plazmové glukózy,
*x*6 (GROUP) značí klinickou skupinu diabetes:
1 je zřetelný, vyložený diabetik,
2 je chemický,
3 je normální zdravý člověk.

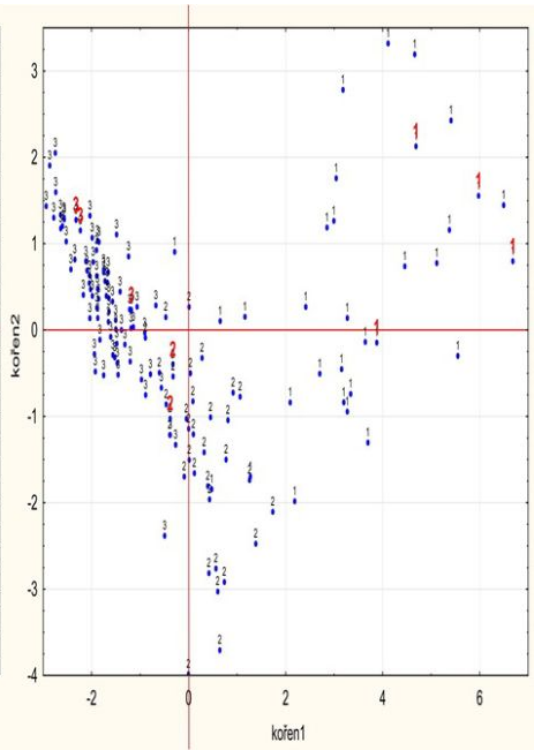
i	*x1*	*x2*	*x3*	*x4*	*x5*	*x6*
1	0.81	80	356	124	55	3
...	...	...	...	...	...	...
144	1.11	328	1246	124	442	1











Úloha B40 *Vliv vlákninové diety ve snídaňových lupínkách na nadváhu žen*

Ženy konzumují lupínky různých typů, např. otrubové lupínky, gumové, obojí a konečně i s kontrolním druhem lupínek bez vlákniny k snídani tolik, na kolik mají hlad. Kontigenční tabulka ukazuje vztah 4 různých druhů lupínek a 4 hladin nevolnosti z přesycení. χ^2-test může otestovat, zda je přesycení nezávislé na druhu lupínek.

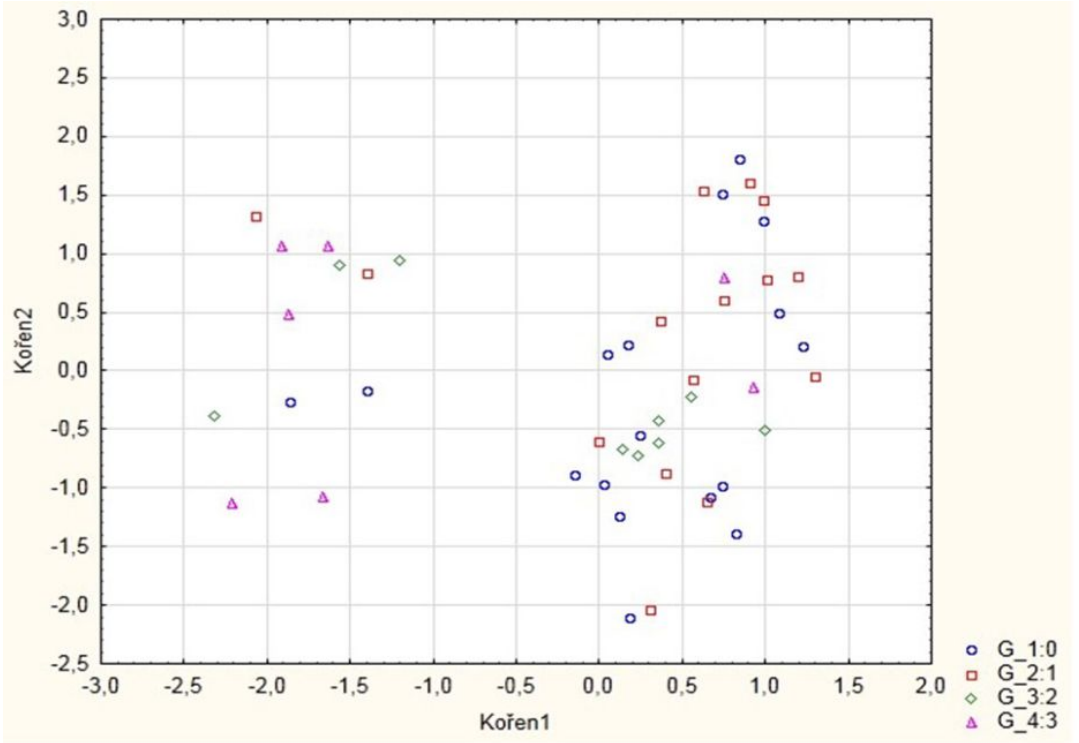
U diskriminační analýzy vezměte za závisle proměnnou údaje o stupni přesycení $x5$ (BLOAT) a ostatní znaky pak za nezávisle diskriminátory.

◯ ***Data:*** Celkem 12 žen v řádcích bylo testováno na 3 diety ve sloupcích a na kontrolní dietu. Před každým jídlem otylky jedly lupínky a) s otrubami, b) s gumovými vlákny, c) s kombinací obou, a konečně d) bez vláken. Celkem bylo provedeno 48 pokusů a užito 5 sledovaných znaků.

Datový výběr CRACKER obsahuje:

i (ID) pořadové číslo ženy,
*x*1 (CRACKER) značící typ vlákna v lupínkách,
*x*2 (DIET) značí jednu ze 4 diet čili typu lupínek,
*x*3 (IDFEM) index sledované ženy,
*x*4 (DIGESTED) značí strávené kalorie.
*x*5 (BLOAT) značí stupeň přesycení a nadýmání po jídle: 0 žádný, 1 malý, 2 střední a 3 vysoký.

i	*x1*	*x2*	*x3*	*x4*	*x5*
1	control	1	3	1772.84	0
	...	...	...	...	...
12	bran	4	12	2287.52	0



Úloha E22 *Zdravotní klasifikace 77 druhů obilninových lupínek k snídani*

Zdravotní norma u lupínek požaduje, že dospělí by měli denně zkonzumovat ne více než 30% kalorií ve formě tuku, muži potřebují okolo 63 g bílkovin a ženy 50 g bílkovin a zbytek poskytnou uhlovodíky (1 cal = 4.184 joulu). 1 g tuku obsahuje asi 9 kalorií, bílkoviny a uhlovodíky 4 kalorie na 1 g. Vhodná dieta by měla obsahovat 20 - 35 g dietní vlákniny. Níže analyzovaná data obsahují počet kalorií na 1 porci, gramy bílkovin a tuků, mg Na^+ a K^+iontů, gramy vlákniny, uhlovodíků a cukru, typické procento FDA RDA vitaminů, hmotnost jedné porce a počet šálků na 1 porci podávanou. Data obsahují umístění na regále v prodejně: nahoře, ve středu či dole u země.

◯ ***Data:*** Výběr CEREALS obsahuje 77 druhů obilninových lupínek v řádcích a 15 znaků:

i (ID) je index druhu lupínek, *j* (J) je název druhu obilninových lupínek,

*x*1 (MFR) značí výrobce lupínek: A je American Home Food Products; G je General Mills; K je Kelloggs; N je Nabisco; P je Post; Q je Quaker Oats; R je Ralston Purina,

*x*2 (TYPE) značí formu v jaké je lupínka konzumována: C je studená, H je horká,

*x*3 (CALORIES) značí počet kalorií jedné porce,

*x*4 (PROTEIN) značí obsah bílkovin v gramech v jedné porci,

*x*5 (FAT) značí obsah tuku v gramech v jedné porci,

*x*6 (SODIUM) značí obsah sodných iontů v miligramech v jedné porci,

*x*7 (FIBER) značí obsah dietní vlákniny v jedné porci,

*x*8 (CH) značí obsah komplexních uhlovodíků v gramech,

*x*9 (SUGARS) značí obsah cukrů v gramech v jedné porci,

*x*10 (POTASS) značí obsah draselných iontů v miligramech v jedné porci,

*x*11 (VITAMINS) značí obsah vitaminů a minerálů v jedné porci (0, 25, nebo 100) ukazuje typické % doporučených FDA,

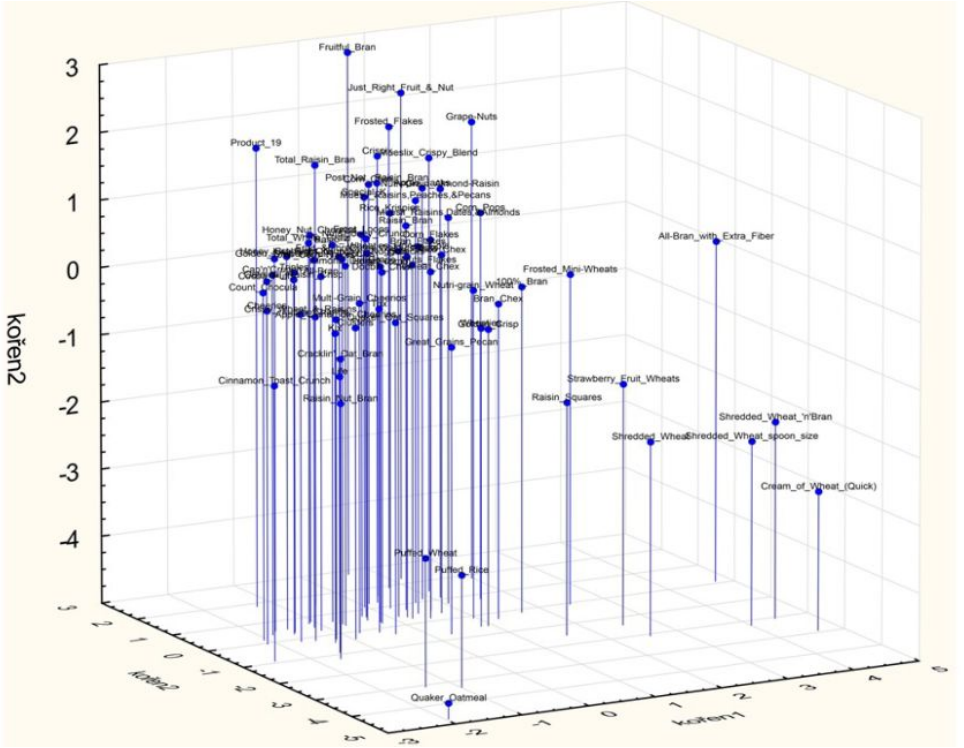
*x*12 (SHELF) značí umístění na regále, jehož výška je číslovaná směrem od země 1 (nejníže), 2, nebo 3 (nejvýše),

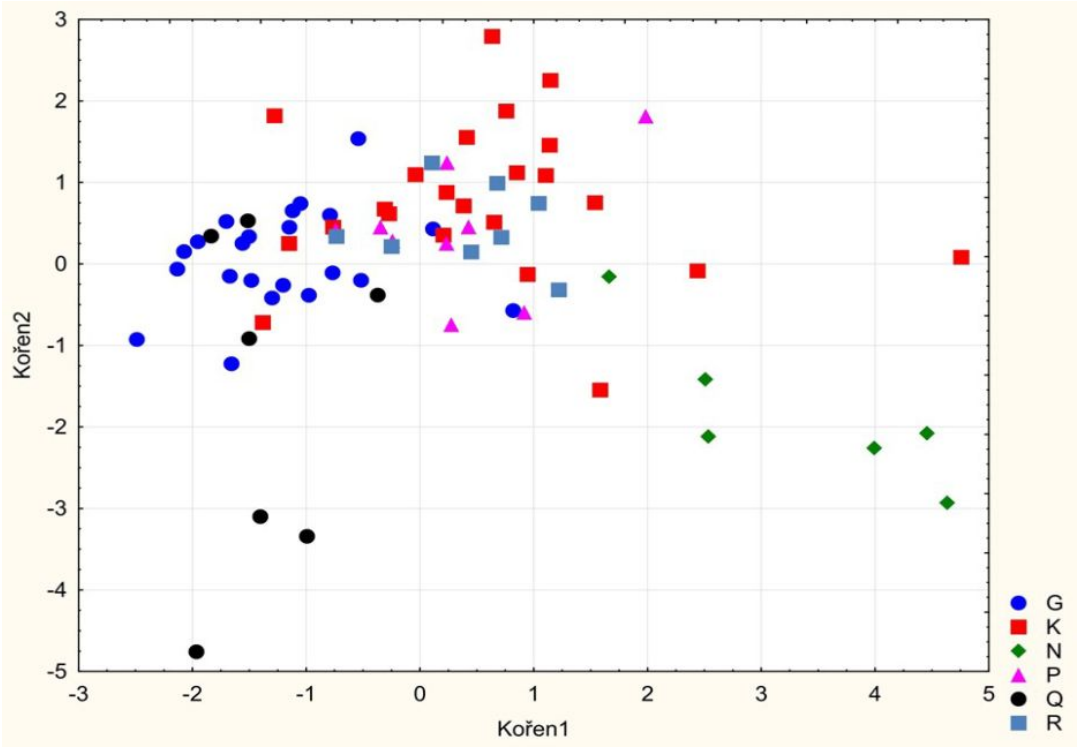
*x*13 (WEIGHT) značí hmotnost jedné porce v uncích,

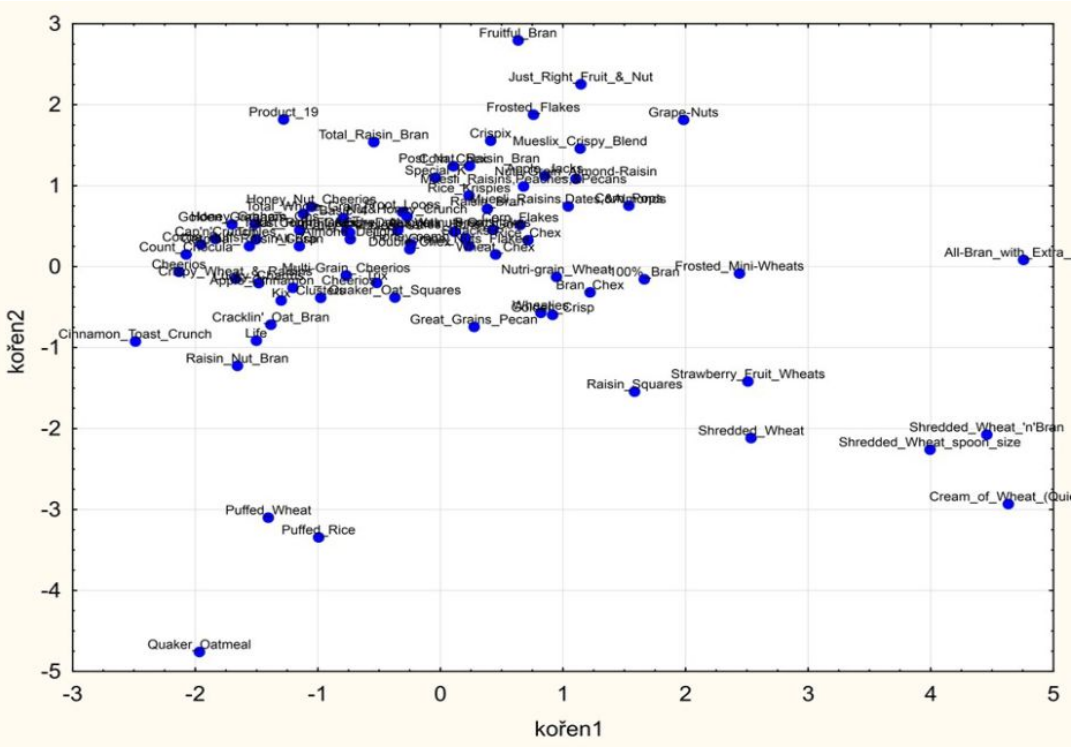
*x*14 (CUPS) značí počet šálků v jedné porci,

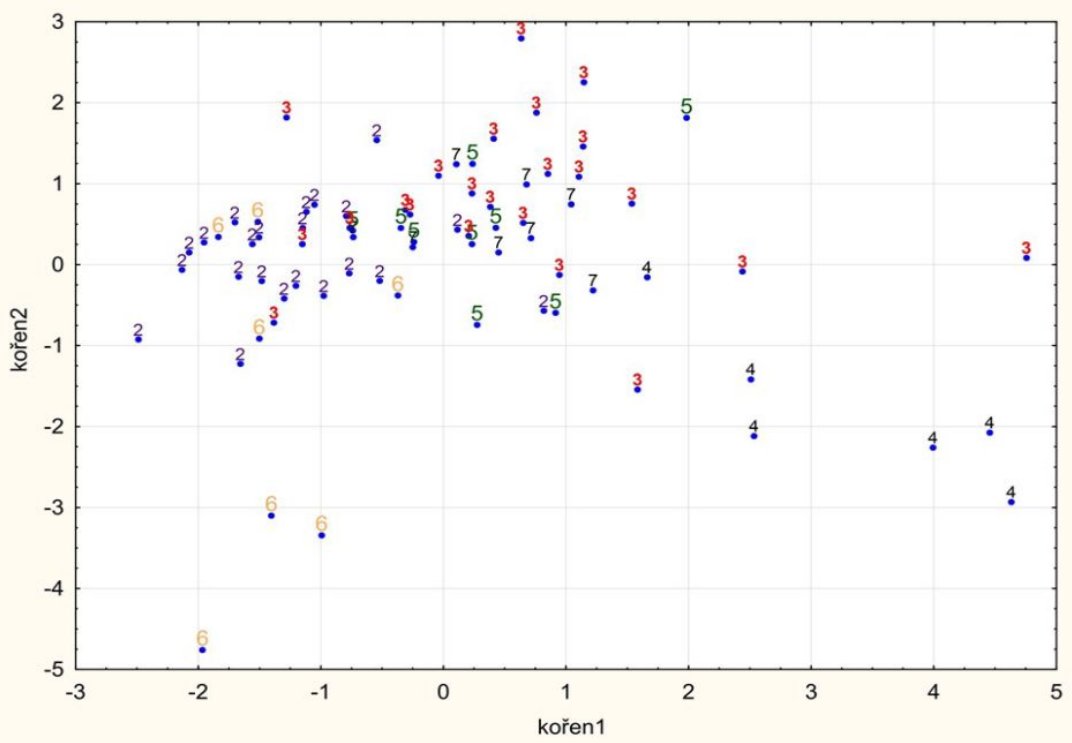
*x*15 (RATING) značí hodnocení, ocenění druhu lupínek.

Číslo -1 značí chybějící numerickou informaci.









Kořeny odstraněny	Test chí-kvadrát po odstranění post. kořenů (E22)					
	Vlastní číslo	Kan. R	Wilk. Lambda	Chi-kv.	sv	p-hodn.
0	1,824023	0,803676	0,100675	149,2308	70	0,000000
1	0,971859	0,702043	0,284308	81,7502	52	0,005270
2	0,428362	0,547629	0,560616	37,6167	36	0,395068
3	0,199115	0,407495	0,800763	14,4424	22	0,885038
4	0,041442	0,199482	0,960207	2,6394	10	0,988685

Proměnná	Klasifikační funkce; grupovací : MFR (E22)				
	G_1:2 p=,28947	G_2:3 p=,30263	G_3:4 p=,07895	G_4:5 p=,11842	G_5:6 p=,10526
TYPE	431	429	443	429	431
CALORIES	1444741	1444724	1444645	1444730	1444725
PROTEIN	-21234981	-21234726	-21233572	-21234813	-21234741
FAT	10976930	10976795	10976200	10976840	10976805
SODIUM	353534	353530	353511	353531	353530
FIBER	-22336642	-22336371	-22335156	-22336463	-22336389
CH	-7087424	-7087338	-7086952	-7087368	-7087344
SUGARS	4703311	4703254	4702999	4703273	4703257
POTASS	220361	220359	220347	220360	220359
VITAMINS	332383	332379	332361	332380	332379
SHELF	1966	1966	1963	1966	1967
WEIGHT	41729	41722	41718	41721	41730
CUPS	-5056	-5058	-5059	-5059	-5054
RATING	6487518	6487440	6487088	6487467	6487445
Konstant	-178184354	-178180076	-178160726	-178181531	-178180305

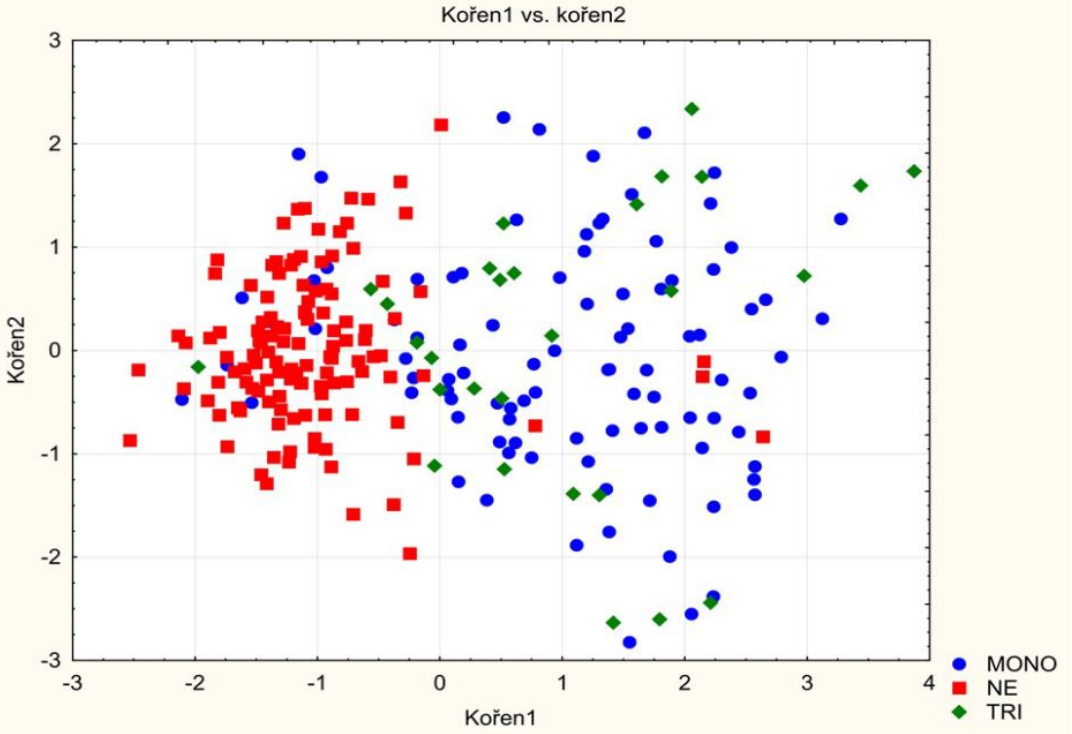
Úloha B77 *Rozlišení mono-, di- a trifazické hormonální antikoncepce u žen*

247 žen se dělí do tří tříd podle užívání hormonální antikoncepce.
V 1. třídě je 123 žen, které neužívají hormonální antikoncepci (**HA** = NE) – zavisle p.,
ve 2. třídě je 152 žen užívající monofazickou (**HA** = MONO) antikoncepci,
ve 3. třídě je 95 žen, které užívají trifazickou hormonální antikoncepci (**HA** = TRI).

Cílem diskriminační analýzy je zjistit, zda lze na základě diskriminátorů zde steroidních hormonů a proteinem přenášejícím steroidní hormony (**SHBG**) rozpoznat ženy neužívající a užívající hormonální antikoncepci, a dále rozpoznat mono- a trifazickou antikoncepci, lišící se dávkováním účinné látky.

◯ *Data:* Zdrojová data s popisem znaků ve sloupcích 5 diskriminátory
(Kortizol, Testosteron, SHBG, DHEA a DHEAS).

HA	Kortizol	SHBG	DHEAS	DHEA	Testosteron
i	x1	x2	x3	x4	x5
MONO	1383	85.2	5.68	46.28	1.62
.....	.....	.....	.....	.....	.....
TRI	620	86.5	4.16	12.02	1.68



Kořeny odstraněny	Test chí-kvadrát po odstranění post. kořenů (B68-B77) Vlastní číslo	Kan. R	Wilk. Lambda	Chi-kv.	sv	p-hodn.
0	1,087462	0,721768	0,474520	180,3991	10	0,000000
1	0,009547	0,097245	0,990543	2,2994	4	0,680877

Skup.	Průměry kan. proměnných (B68-B77) Kořen1	Kořen2
MONO	1,02516	-0,076571
NE	-1,04063	0,000854
TRI	1,05541	0,247217

Proměnná	Standardiz. koeficienty (B68-B77) pro kanonické proměnné Kořen1	Kořen2
B77x1	0,545676	0,915162
B77x2	0,723332	-0,865965
B77x3	0,204730	-0,234211
B77x4	-0,277455	-0,206162
B77x5	-0,072968	0,261355
Vlastní	1,087462	0,009547
KumPodíl	0,991297	1,000000

Proměnná	Faktorová strukturní matice (B68-B77) Korelační proměnné - Kanonické kořeny (vnitřní korelace) Kořen1	Kořen2
B77x1	0,705334	0,637291
B77x2	0,840542	-0,458747
B77x3	-0,162699	0,016145
B77x4	-0,157594	0,037997
B77x5	0,045107	0,119113

Úloha E19 *Klasifikace smrku ztepilého z pokusné plochy v Orlických horách*

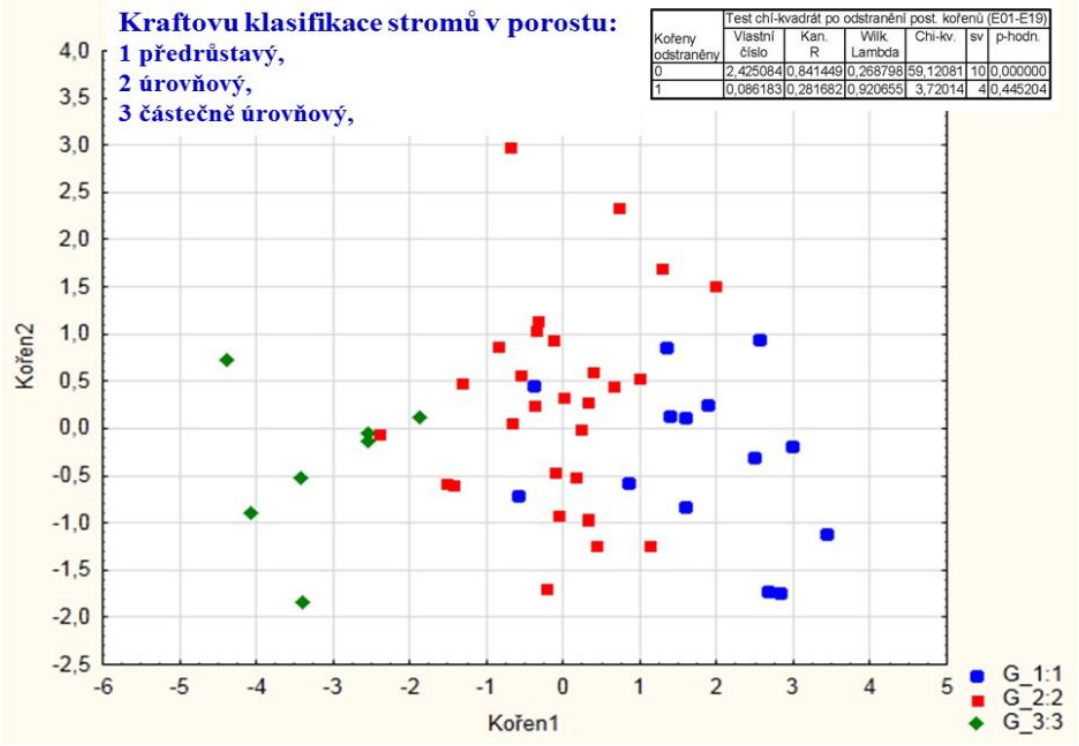
Při klasifikaci 50 stromů smrku ztepilého (Picea abies L.) z pokusné plochy v Orlických horách byl vyšetřován dle 6 charakteristik každý strom zvlášť.

Data: Výběr STROMY:

i (ID) index stromu,
*x*1 (D) značí výčetní tloušťku stromu v cm, měřenou ve výšce 1.3 m nad zemí,
*x*2 (RTP) značí roční tloušťkový přírůst v cm (tamtéž),
*x*3 (H) značí výšku stromu v m,
***x*4 (KRAFT) značí Kraftovu klasifikaci stromů v porostu: 1 předrůstavý, 2 úrovňový, 3 částečně úrovňový,**
*x*5 (ZDRAV) značí zdravotní stav: 1 olistění 90 – 100 %, 2 olistění 70 – 80 %, 3 olistění pod 60 %),
*x*6 (RAŠ) značí rašení letorostů: 1 pozdní, 2 střední, 3 časné.

i	*x1*	*x2*	*x3*	*x4*	*x5*	*x6*
1	8.9	0.8	4.60	2	1	2
...	...	...	...	...	...	...
50	7.0	0.5	5.80	2	3	3

Kraftovu klasifikace stromů v porostu:
1 předrůstavý,
2 úrovňový,
3 částečně úrovňový,



Kořeny odstraněny	Test chí-kvadrát po odstranění post. kořenů (E01-E19) Vlastní číslo	Kan. R	Wilk. Lambda	Chi-kv.	sv	p-hodn.
0	2,425084	0,841449	0,268798	59,12081	10	0,000000
1	0,086183	0,281682	0,920655	3,72014	4	0,445204

Kořeny odstraněny	Test chí-kvadrát po odstranění post. kořenů (E01-E19) Vlastní číslo	Kan. R	Wilk. Lambda	Chi-kv.	sv	p-hodn.
0	2,425084	0,841449	0,268798	59,12081	10	0,000000
1	0,086183	0,281682	0,920655	3,72014	4	0,445204

Skup.	Průměry kan. proměnných (E01-E19) Kořen1	Kořen2
G_1:1	1,75688	-0,314043
G_2:2	-0,08173	0,241716
G_3:3	-3,17518	-0,373307

Proměnná	Faktorová strukturní matice (E01-E19) Korelační proměnné - Kanonické kořeny (vnitřní korelace) Kořen1	Kořen2
E19x1	0,806638	0,521042
E19x2	0,277166	0,507606
E19x3	0,653265	-0,292438
E19x5	0,272048	-0,523211
E19x6	0,033093	0,174083

Proměnná	Klasifikační funkce; grupovací : E19x4 (E01-E19) G_1:1 p=,28000	G_2:2 p=,58000	G_3:3 p=,14000
E19x1	9,1286	7,8943	4,7326
E19x2	-12,7102	-9,8558	-7,5089
E19x3	10,0572	8,4413	7,5413
E19x5	7,4350	5,8572	4,0769
E19x6	2,0369	2,3188	2,1130
Konstant	-84,6280	-62,2406	-35,8244

Skup.	Klasifikační matice (E01-E19) Řádky: pozorované klasifikace Sloupce: předpovězené klasifikace % správnýc	G_1:1 p=,28000	G_2:2 p=,58000	G_3:3 p=,14000
G_1:1	71,42857	10	4	0
G_2:2	89,65517	2	26	1
G_3:3	85,71429	0	1	6
Celkem	84,00000	12	31	7

Proměnná	Standardiz. koeficienty (E01-E19) pro kanonické proměnné Kořen1	Kořen2
E19x1	0,857251	0,679610
E19x2	-0,197606	0,291570
E19x3	0,369694	-0,827552
E19x5	0,449677	-0,368888
E19x6	-0,017054	0,361201
Vlastní	2,425084	0,086183
KumPodíl	0,965681	1,000000