Téma č. 5.: Lineární diskriminační analýza pro dvě skupiny

Příklad: Třídění lebek Tibeťanů (Příklad je převzat z knihy Meloun M., Militký J., Hill, M.: Počítačová analýza vícerozměrných dat v příkladech. Academia Praha 2005)
Datový soubor lebky.sta obsahuje údaje o 32 lebkách nalezených na pohřebištích v Tibetu.
Sledují se tyto proměnné:
ID … identifikátor (1 pro lebky z okolí Sikkimu, 2 pro lebky z okolí Lhasy)
Ldelka … největší délka lebky (v mm)
Lsirka … největší horizontální šířka lebky (v mm)
Lvyska … výška lebky (v mm)
Ovyska … výška horní části obličeje (v mm)
Osirka … šířka obličeje mezi body lícních kostí (v mm)

	1 ID	2 Ldelka	3 Lsirka	4 Lvyska	5 Ovyska	6 Osirka
1	1	190,5	152,5	145,0	73,5	136,5
2	1	172,5	132,0	125,5	63,0	121,0
3	1	167,0	130,0	125,5	69,5	119,5
4	1	169,5	150,5	133,5	64,5	128,0
5	1	175,0	138,5	126,0	77,5	135,5
6	1	177,5	142,5	142,5	71,5	131,0
7	1	179,5	142,5	127,5	70,5	134,5
8	1	179,5	138,0	133,5	73,5	132,5
9	1	173,5	135,5	130,5	70,0	133,5
10	1	162,5	139,0	131,0	62,0	126,0
11	1	178,5	135,0	136,0	71,0	124,0
12	1	171,5	148,5	132,5	65,0	146,5
13	1	180,5	139,0	132,0	74,5	134,5
14	2	183,0	149,0	121,5	76,5	142,0
15	2	169,5	130,0	131,0	68,0	119,0
16	2	172,0	140,0	136,0	70,5	133,5
17	2	170,0	126,5	134,5	66,0	118,5
18	2	182,5	136,0	138,5	76,0	134,0
19	2	179,5	135,0	128,5	74,0	132,0
20	2	191,0	140,5	140,5	72,5	131,5
21	2	184,5	141,5	134,5	76,5	141,5
22	2	181,0	142,0	132,5	79,0	136,5
23	2	173,5	136,5	126,0	71,5	136,5
24	2	188,5	130,0	143,0	79,5	136,0
25	2	175,0	153,0	130,0	76,5	142,0
26	2	196,0	142,5	123,5	76,0	134,0
27	2	200,0	139,5	143,5	82,5	146,0
28	2	185,0	134,5	140,0	81,5	137,0
29	2	174,5	143,5	132,5	74,0	136,5
30	2	195,5	144,0	138,5	78,5	144,0
31	2	197,0	131,5	135,0	80,5	139,0
32	2	182,5	131,0	135,0	68,5	136,0

Úkolem je najít Fisherovu lineární diskriminační funkci, která pomocí proměnných Ldelka až Osirka umožní rozlišit lebky z okolí Sikkimu od lebek z okolí Lhasy.

Výsledky (s částečným návodem)

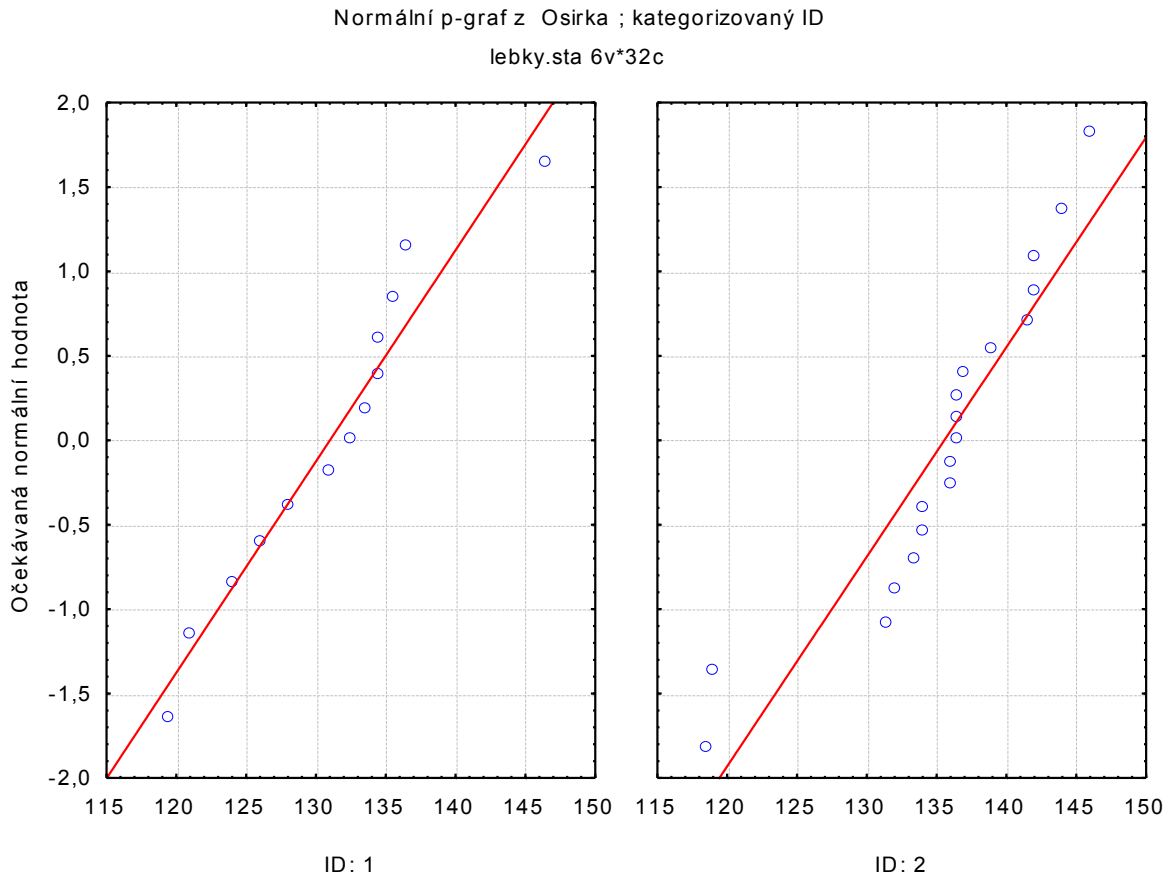
Testování hypotézy o normalitě sledovaných proměnných v daných dvou skupinách pomocí Lilieforsovy varianty K-S testu a pomocí S-W testu:

Proměnná	Testy normality (lebky.sta) Zhrnout podmínku: ID=1				
	N	max D	Lilliefors p	W	p
Ldelka	13	0,149568	p > .20	0,971258	0,908822
Lsirka	13	0,188580	p > .20	0,946284	0,543198
Lvyska	13	0,196329	p < ,20	0,900168	0,134439
Ovyska	13	0,176191	p > .20	0,944919	0,523649
Osirka	13	0,148589	p > .20	0,954446	0,666891

Proměnná	Testy normality (lebky.sta) Zhrnout podmínku: ID=2				
	N	max D	Lilliefors p	W	p
Ldelka	19	0,121226	p > .20	0,946640	0,345905
Lsirka	19	0,099534	p > .20	0,973572	0,844925
Lvyska	19	0,116289	p > .20	0,969669	0,769812
Ovyska	19	0,149966	p > .20	0,965452	0,683230
Osirka	19	0,180030	p < ,10	0,873328	0,016463

Vidíme, že ve 2. skupině zamítá S-W test hypotézu o normalitě proměnné Osirka na hladině významnosti 0,05, Lilieforsův test nikoli.

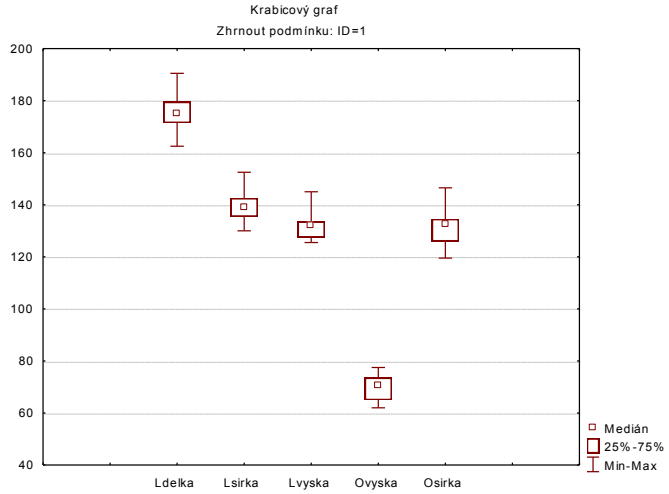
N-P plot pro proměnnou Osirka v 1. a 2. skupině



Odhad vektorů středních hodnot v 1. skupině:

	Popisné statistiky (lebky.sta) Zhrnout podmínku: ID=1
Proměnná	Průměr
Ldelka	175,1923
Lsirka	140,2692
Lvyska	132,3846
Ovyska	69,6923
Osirka	131,0000

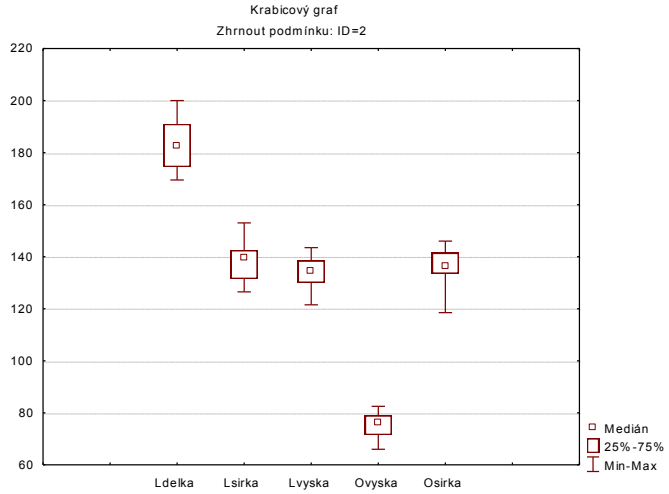
Krabicové grafy všech proměnných v 1. skupině:



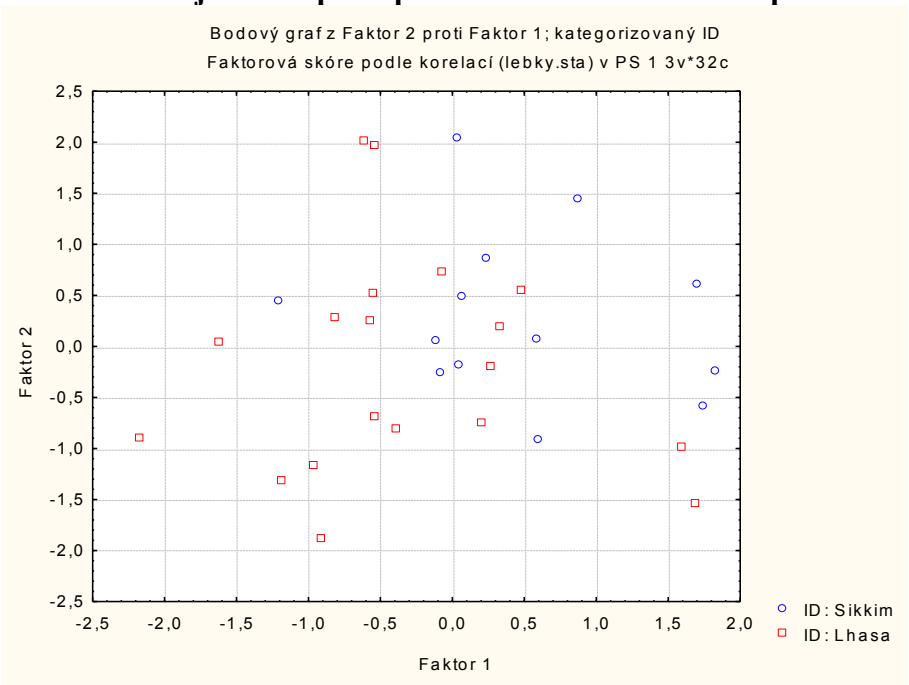
Odhad vektorů středních hodnot ve 2. skupině:

	Popisné statistiky (lebky.sta) Zhrnout podmínku: ID=2
Proměnná	Průměr
Ldelka	183,1842
Lsirka	138,2368
Lvyska	133,9211
Ovyska	75,1579
Osirka	135,5526

Krabicové grafy všech proměnných ve 2. skupině:



Rozmístění objektů na ploše prvních dvou hlavních komponent:



Odhad varianční matice v 1. skupině:

	Kovariance (lebky.sta) Zhrnout podmínku: ID=1				
Proměnná	Ldelka	Lsirka	Lvyska	Ovyska	Osirka
Ldelka	50,02244	17,52724	25,02404	22,85577	20,04167
Lsirka	17,52724	47,31731	25,57532	-0,76442	32,35417
Lvyska	25,02404	25,57532	36,83974	5,89904	12,27083
Ovyska	22,85577	-0,76442	5,89904	22,64744	10,29167
Osirka	20,04167	32,35417	12,27083	10,29167	53,25000

Odhad varianční matice ve 2. skupině:

	Kovariance (lebky.sta) Zhrnout podmínku: ID=2				
Proměnná	Ldelka	Lsirka	Lvyska	Ovyska	Osirka
Ldelka	90,31140	7,3012	20,1126	31,64985	39,07310
Lsirka	7,30117	47,2880	-14,6747	10,68275	30,51462
Lvyska	20,11257	-14,6747	38,1462	8,66594	4,42105
Ovyska	31,64985	10,6827	8,6659	22,14035	25,13012
Osirka	39,07310	30,5146	4,4211	25,13012	51,05263

Odhad společné varianční matice (výpočet podle vzorce $\mathbf{S} = \frac{(n_1 - 1)\hat{\mathbf{S}}_1 + (n_2 - 1)\hat{\mathbf{S}}_2}{n_1 + n_2 - 2}$):

Proměnná	Ldelka	Lsirka	Lvyska	Ovyska	Osirka
Ldelka	74,19582	11,3916	22,07716	28,13222	31,46053
Lsirka	11,3916	47,29973	1,425304	6,10388	31,25044
Lvyska	22,07716	1,425304	37,62362	7,559177	7,560965
Ovyska	28,13222	6,10388	7,559177	22,34318	19,19474
Osirka	31,46053	31,25044	7,560965	19,19474	51,93158

Odhad společné varianční matice (pomocí SPSS):

Pooled Within-Groups Matrices [a]

		Ldelka	Lsirka	Lvyska	Ovyska	Osirka
Covariance	Ldelka	74,196	11,392	22,077	28,132	31,461
	Lsirka	11,392	47,300	1,425	6,104	31,250
	Lvyska	22,077	1,425	37,624	7,559	7,561
	Ovyska	28,132	6,104	7,559	22,343	19,195
	Osirka	31,461	31,250	7,561	19,195	51,932

a. The covariance matrix has 30 degrees of freedom.

Boxův test shody variančních matic (pomocí SPSS):

Test Results

Box's M		22,653
F	Approx.	1,218
	df 1	15
	df 2	2651,539
	Sig.	,249

Tests null hypothesis of equal population covariance matrices.

Hypotézu o shodě variančních matic nezamítáme na asymptotické hladině významnosti 0,05, protože p-hodnota = 0,249 je větší než 0,05.

Test shody vektorů středních hodnot:
Statistiky – Základní statistiky/tabulky – t-test, nezávislé, dle skupin – OK – Proměnné – Závisle proměnné Ldelka až Osirka, Grupovací proměnná ID – OK – na záložce Možnosti zaškrtneme Vícerozměrný test – Výpočet

t-testy; grupováno: ID (lebky.sta)
Skup. 1: 1; Skup. 2: 2
T2(celé případy 14,0638 F(5,26)=2,4377 p<,06127

Proměnná	Průměr 1	Průměr 2	t	sv	p	Poč.plat 1	Poč.plat. 2	Sm.odch. 1	Sm.odch. 2	F-poměr Rozptyly	p Rozptyly
Ldelka	175,1923	183,1842	-2,57771	30	0,015102	13	19	7,072654	9,503231	1,805418	0,298973
Lsirka	140,2692	138,2368	0,82101	30	0,418115	13	19	6,878758	6,876628	1,000620	0,970788
Lvyska	132,3846	133,9211	-0,69592	30	0,491837	13	19	6,069575	6,176261	1,035463	0,976491
Ovyska	69,6923	75,1579	-3,21246	30	0,003136	13	19	4,758932	4,705353	1,022903	0,938035
Osirka	131,0000	135,5526	-1,75517	30	0,089438	13	19	7,297260	7,145112	1,043041	0,909092

Testová statistika se realizuje hodnotou 2,4377, odpovídající p.hodnota je menší než 0,06127, tedy na hladině významnosti 0,056 nezamítáme hypotézu o shodě vektorů středních hodnot. Individuální t-testy však prokázaly, že na hladině významnosti 0,05 se liší střední hodnoty proměnných Ldelka a Ovyska.

Stanovení odhadů apriorních pravděpodobností:

$$p_1 = \frac{n_1}{n} = \frac{13}{32} = 0{,}40625\,,\; p_2 = \frac{n_2}{n} = \frac{19}{32} = 0{,}59375$$

Stanovení odhadu Fisherovy lineární diskriminační funkce:
Statistiky – Vícerozměrné průzkumné techniky – Diskriminační analýza - Pro-měnné – Grupovací proměnná ID, Seznam nezávislých proměnných Ldelka až Osirka – OK – OK – na záložce Klasifikace zvolíme Klasifikační funkce. Do výstupní tabulky přidáme novou proměnnou, do jejíhož Dlouhého jména napíšeme =v1-v2

	Klasifikační funkce; grupovací : ID (lebky.sta)		
Proměnná	G_1:1 p=,40625	G_2:2 p=,59375	NProm =v1-v2
Ldelka	1,168	1,202	-0,03346
Lsirka	2,820	2,692	0,128157
Lvyska	2,748	2,722	0,025791
Ovyska	0,280	0,454	-0,17415
Osirka	-0,385	-0,302	-0,0839
Konstant	-467,373	-475,503	8,130393

Posouzení účinnosti diskriminace resubstituční metodou:
Na záložce Klasifikace zvolíme Klasifikační matice.

	Klasifikační matice (lebky.sta) Řádky: pozorované klasifikace Sloupce: předpovězené klasifikace		
Skup.	% správnýc	Sikkim p=,40625	Lhasa p=,59375
Sikkim	69,23077	9	4
Lhasa	84,21053	3	16
Celkem	78,12500	12	20

Pro určení chybně zařazených případů zvolíme na záložce Klasifikace možnost Klasifikace případů. Zjistíme, že v 1. skupině došlo k mylnému zařazení u lebek č. 5, 8, 9 a 13, ve 2. skupině u lebek číslo 15, 16, 17.

Porovnání s náhodnou klasifikací:
Odhad celkové pravděpodobnosti mylné klasifikace je

$$2p_1(1-p_1) = 2 \cdot \frac{13}{32} \cdot \frac{19}{32} = 0{,}4824.$$

Použitím diskriminační analýzy jsme tedy dosáhli značného zlepšení, pravděpodobnost mylné klasifikace klesla na 0,22.