Přednáška č. 1.: Tabulkové a grafické zpracování vícerozměrných dat

Osnova

1. Tabulkové zpracování

a) Kontingenční tabulky, statistická indukce pro KT
b) Tabulky číselných charakteristik, statistická indukce (dvouvýběrový t-test a jeho neparametrické obdoby, jednofatorová ANOVA a její neparametrické obdoby)
c) Asociační tabulky (korelační matice, matice vzdáleností)

2. Grafické zpracování

a) 3D sloupkové diagramy
b) Vícenásobné krabicové diagramy
c) Dvourozměrné tečkové diagramy
d) Bag plot
e) Ikonové grafy

Motivace: Při statistickém zpracování dat se často setkáváme s vícerozměrnými daty. Vyskytují se v situacích, kdy u každého z n objektů zjišťujeme hodnoty p znaků, které označíme $X_1, ..., X_p$. Dostáváme tak p-rozměrný datový soubor ve formě matice n x p:

$$\begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \cdots & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix}.$$

Řádky této matice se vztahují k jednotlivým objektům, zatímco sloupce k jednotlivým znakům. Prvotní informace o datech můžeme získat tabulkovou nebo grafickou formou.

Příklad: Máme k dispozici následující údaje o 32 lidech:
proměnná X_1 (Sex) ……..… udává pohlaví (1 muž, 2 žena)
proměnná X_2 (Vlasy) …….. udává stav vlasů (0 málo nebo žádné; 1 dost)
proměnná X_3 (Věk) ……. udává věk v počtu dovršených let
proměnná X_4 (IQ)…….... udává hodnotu IQ
proměnná X_5 (Výška) …..... udává výšku v cm
proměnná X_6 (Hmotnost) … udává hmotnost v kg
proměnná X_7 (Boty) ……..... udává velikost obuvi (v evropském číslování)
proměnná X_8 (Příjem) ...…... udává měsíční příjem v korunách
proměnná X_9 (Pivo) ……. udává počet vypitých litrů piva za rok
proměnná X_{10} (Víno) …….. udává počet vypitých litrů vína za rok
Určete typy znaků.
Řešení: X_1, X_2 – nominální znaky (alternativní, nabývají pouze dvou variant), X_3, X_5, …, X_{10} – poměrové znaky, X_4 … intervalový znak.

1. Tabulkové zpracování

a) Kontingenční tabulky

Nechť znaky X_i a X_j jsou nominálního typu. Označme znak X_i jako X a znak X_j jako Y. Předpokládáme, že znak X má r variant a znak Y má s variant. V daném dvourozměrném datovém souboru zjistíme simultánní absolutní četnosti n_{jk} dvojic variant $(x_{[j]}, y_{[k]})$ a zapíšeme je do kontingenční tabulky:

	y	$y_{[1]}$	...	$y_{[s]}$	$n_{j.}$
x	n_{jk}				
$x_{[1]}$		n_{11}	...	n_{1s}	$n_{1.}$
⋮		...	...	...	...
$x_{[r]}$		n_{r1}	...	n_{rs}	$n_{r.}$
$n_{.k}$		$n_{.1}$	...	$n_{.s}$	n

$n_{j.} = n_{j1} + ... + n_{js}$ – marginální absolutní četnost varianty $x_{[j]}$
$n_{.k} = n_{1k} + ... + n_{rk}$ – marginální absolutní četnost varianty $y_{[k]}$
Dále můžeme vypočítat sloupcově a řádkově podmíněné relativní četnosti:

$p_{j(k)} = \frac{n_{jk}}{n_{.k}}$ - sloupcově podmíněná relativní četnost varianty $x_{[j]}$ za předpokladu $y_{[k]}$

$p_{(j)k} = \frac{n_{jk}}{n_{j.}}$ - řádkově podmíněná relativní četnost varianty $y_{[k]}$ za předpokladu $x_{[j]}$.

Statistická indukce pro kontingenční tabulky: viz přednáška č. 12 předmětu Aplikovaná statistika 1.

Příklad: Pro proměnné X_1 (Sex) a X_2 (Vlasy) vytvořte kontingenční tabulku simultánních absolutních četností a sloupcově a řádkově podmíněných relativních četností. Na hladině významnosti 0,05 testujte pomocí Fisherova přesného testu hypotézu, že proměnné Sex a Vlasy jsou nezávislé. Vypočtěte také Cramérův koeficient.

Řešení pomocí systému STATISTICA:

KT simultánních absolutních četností

Sex	Vlasy malo	Vlasy dost	Řádk. součty
muz	15	1	16
zena	1	15	16
Vš.skup.	16	16	32

Ve výběrovém souboru bylo 16 mužů a 16 žen. 15 mužů má málo vlasů a jeden má vlasů dost. U žen je tomu přesně naopak.

KT sloupcově podmíněných relativních četností:

	Sex	Vlasy malo	Vlasy dost	Řádk. součty
Četnost	muz	15	1	16
Sloupc. četn.		93,75%	6,25%	
Četnost	zena	1	15	16
Sloupc. četn.		6,25%	93,75%	
Četnost	Vš.skup.	16	16	32

Z osob, které mají málo vlasů, je 93,75% mužů a 6,25% žen. Z osob, které mají vlasů dost, je 6,25% mužů a 93,75% žen.

KT řádkově podmíněných relativních četností:

	Sex	Vlasy malo	Vlasy dost	Řádk. součty
Četnost	muz	15	1	16
Řádk. četn.		93,75%	6,25%	
Četnost	zena	1	15	16
Řádk. četn.		6,25%	93,75%	
Četnost	Vš.skup.	16	16	32

Z mužů má málo vlasů 93,75% a dost vlasů 6,25%. Z žen má málo vlasů 6,25% a dost vlasů 93,75%.

Výstupní tabulka Fisherova testu:

Statist.	Chí-kvadr.	sv	p
Pearsonův chí-kv.	24,50000	df=1	p=,00000
M-V chí-kvadr.	29,39875	df=1	p=,00000
Yatesův chí-kv.	21,12500	df=1	p=,00000
Fisherův přesný, 1-str.			p=,00000
2-stranný			p=,00000
McNemarův chí-kv. (A/D)	,0333333	df=1	p=,85513
(B/C)	,5000000	df=1	p=,47950

p-hodnota Fisherova testu je blízká 0, je mnohem menší než hladina významosti 0,05, tedy hypotézu o nezávislosti proměnných Sex a Vlasy zamítáme na hladině významnosti 0,05.

Výpočet Cramérova koeficientu:

	Statist. : Sex(2) x Vlasy(2) (Lide.sta)		
Statist.	Chí-kvadr.	sv	p
Pearsonův chí-kv.	24,50000	df=1	p=,00000
M-V chí-kvadr.	29,39875	df=1	p=,00000
Fí pro tabulky 2 x 2	,8750000		
Tetrachorická korelace	,9811733		
Kontingenční koeficient	,6585046		

Cramérův koeficient je zde označen symbolem Fí. Nabývá hodnoty 0,875, tedy mezi proměnnými Sex a Vlasy existuje silná závislost.

Řešení pomocí systému SPSS:

Vytvoření kontingenční tabulky simultánních absolutních četností a sloupcově a řádkově podmíněných relativních četností:
Analyze – Descriptive Statistics – Crosstabs – Row(s) sex, Column(s) vlasy – Cells - zaškrtneme Percentages Row, Column – Continue – OK. Dostaneme tabulku:

sex * vlasy Crosstabulation

			vlasy		
			málo	dost	Total
sex	muž	Count	15	1	16
		% within sex	93,8%	6,2%	100,0%
		% within vlasy	93,8%	6,2%	50,0%
	žena	Count	1	15	16
		% within sex	6,2%	93,8%	100,0%
		% within vlasy	6,2%	93,8%	50,0%
	Total	Count	16	16	32
		% within sex	50,0%	50,0%	100,0%
		% within vlasy	100,0%	100,0%	100,0%

Interpretace je stejná jako u řešení pomocí systému STATISTICA.

Provedení Fisherova přesného testu a výpočet Cramérova koeficientu:
Analyze – Descriptive Statistics – Crosstabs – Row(s) sex, Column(s) vlasy – zaškrtneme Suppress tables – Statistics – zaškrtneme Phi and Cramer's V – Continue – Exact - zaškrtneme Exact – Continue – OK.

Symmetric Measures

		Value	Approx. Sig.	Exact Sig.
Nominal by Nominal	Phi	,875	,000	,000
	Cramer's V	,875	,000	,000
	N of Valid Cases	32		

p-hodnota Fisherova přesného testu je ve sloupci označeném Exact Sig. Je blízká 0, tedy hypotézu o nezávislosti proměnných Sex a Vlasy zamítáme na hladině významnosti 0,05.

b) Tabulky číselných charakteristik

Nechť znak X_i je nominálního typu a znak X_j je aspoň ordinálního typu. Označme znak X_i jako A a předpokládejme, že má r variant (úrovní). Znak X_j označme jako X. Objekty rozdělíme do r podsouborů podle variant znaku A a v každém podsouboru vypočítáme číselné charakteristiky znaku X (pro intervalový či poměrový znak průměry a směrodatné odchylky, pro ordinální znak mediány).

č. souboru	rozsah	průměr	medián	směrodatná odchylka
1	n_1	m_1	$x1_{0,50}$	s_1
2	n_2	m_2	$x2_{0,50}$	s_2
⋮	⋮	⋮	⋮	⋮
r	n_r	m_r	$xr_{0,50}$	s_r
celkem	n	m	$x_{0,50}$	s

Statistická indukce: pro intervalovou či poměrovou proměnnou X, která se v jednotlivých podsouborech řídí aspoň přibližně normálním rozložením a má v těchto podsouborech shodné rozptyly, se používá jednofaktorová ANOVA (viz přednáška č. 10), v ostatních případech neparametrické testy, např. K-W test či mediánový test (viz přednáška č. 11). Má-li faktor A jen dvě úrovně, lze použít dvouvýběrový t-test (viz přednáška č. 8) nebo dvouvýběrový Wilcoxonův test (viz přednáška č. 11). Dvouvýběrový t-test doplňujeme výpočtem Cohenova koeficientu věcného účinku, který slouží k posouzení vlivu faktoru A na variabilitu hodnot závisle proměnné X.

Příklad: Vytvořte tabulku číselných charakteristik proměnné Příjem rozdělené do dvou skupin podle proměnné Sex. Na hladině významnosti 0,05 testujte hypotézu, že střední hodnoty proměnné Příjem jsou stejné pro muže a ženy. Vypočtěte Cohenův koeficient věcného účinku.

Řešení pomocí systému STATISTICA:

Tabulka číselných charakteristik

Sex	Prijem průměr	Prijem N	Prijem Sm.odch.	Prijem 25.kvan.	Prijem medián	Prijem 75.kvan.
muz	30281,25	16	9117,691	21500,00	32000,00	36500,00
zena	24593,75	16	8025,415	19000,00	24750,00	31750,00
Vš.skup.	27437,50	32	8929,608	19500,00	30000,00	34000,00

Vidíme, že průměrný příjem žen je téměř o 6000 Kč nižší než průměrný příjem mužů. Směrodatná odchylka příjmu žen je o více než 1000 Kč nižší než směrodatná odchylka příjmu mužů. Aspoň čtvrtina žen má příjem nanejvýš 19 000 Kč. Aspoň čtvrtina mužů má příjem aspoň 36 500 Kč.

Výsledky dvouvýběrového t-testu (normalita proměnné Příjem ve skupině mužů a žen byla ověřena S-W testem a na hladině významnosti 0,05 se hypotéza o normalitě nezamítá)

Proměnná	Průměr muz	Průměr zena	t	sv	p	Poč.plat muz	Poč.plat. zena	Sm.odch. muz	Sm.odch. zena	F-poměr Rozptyly	p Rozptyly
Prijem	30281,25	24593,75	1,872954	30	0,070849	16	16	9117,691	8025,415	1,290728	0,627395

Hypotéza o shodě rozptylů se na hladině významnosti 0,05 nezamítá a hypotéza o shodě středních hodnot se na hladině významnosti 0,05 také nezamítá.

Výpočet Cohenova koeficientu $d = \frac{|m_1 - m_2|}{s}$

	1 n1	2 n2	3 m1	4 m2	5 s1	6 s2	7 d
1	16	16	30281,25	24593,75	9117,691	8025,415	0,662189

Hodnota d	účinek
aspoň 0,8	velký
mezi 0,5 až 0,8	střední
mezi 0,2 až 0,5	malý
pod 0,2	zanedbatelný

V našem případě lze považovat vliv pohlaví na variabiltu příjmu za středně velký, avšak na hladině významnosti 0,05 za neprokazatelný.

Řešení pomocí systému SPSS:

Vytvoření tabulky číselných charakteristik proměnné Příjem rozdělené do dvou skupin podle proměnné Sex:
Analyze – Descriptive Statistics – Explore – Dependent List prijem, Factor List sex - zaškrtneme Display Statistics – Statistics – zaškrtneme Descriptives – Continue – OK

Descriptives

sex			Statistic	Std. Error
prijem muž	Mean		30281,25	2279,423
	95% Confidence Interval for Mean	Lower Bound	25422,78	
		Upper Bound	35139,72	
	5% Trimmed Mean		30256,94	
	Median		32000,00	
	Variance		8,313E7	
	Std. Deviation		9117,691	
	Minimum		16000	

		Maximum	45000	
		Range	29000	
		Interquartile Range	16500	
		Skewness	-,316	,564
		Kurtosis	-,950	1,091
	žena	Mean	24593,75	2006,354
		95% Confidence Interval for Mean Lower Bound	20317,31	
		95% Confidence Interval for Mean Upper Bound	28870,19	
		5% Trimmed Mean	24826,39	
		Median	24750,00	
		Variance	6,441E7	
		Std. Deviation	8025,415	
		Minimum	11000	
		Maximum	34000	
		Range	23000	
		Interquartile Range	13375	
		Skewness	-,415	,564
		Kurtosis	-1,088	1,091

Na rozdíl od systému STATISTICA zde uživatel nemůže volit, které číselné charakteristiky ho zajímají a dostane jich tedy mnohem více.

Provedení dvouvýběrového t-testu:
Analyze – Compare Means – Independent-Samples T-test – Test Variable(s) prijem, Grouping Variable sex, Define Groups 1, 2 – Continue - OK

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
prijem	Equal variances assumed	,131	,720	1,873	30	,071	5687,500	3036,647	-514,160	11889,160
	Equal variances not assumed			1,873	29,524	,071	5687,500	3036,647	-518,352	11893,352

Nejprve se podíváme na výsledek Levenova testu homogenity rozptylů. Testová statistika se realizuje hodnotou 0,131, odpovídající p-hodnota je 0,72, tedy hypotézu o shodě rozptylů nezamítáme na hladině významnosti 0,05. Výsledek dvouvýběrového t-testu je tudíž na řádku označeném Equal variances assumed. Testová statistika se realizuje hodnotou 1,873, odpovídající p-hodnota je 0,071, tedy hypotézu o shodě středních hodnot proměnné prijem ve skupině mužů a žen nezamítáme na hladině významnosti 0,05.

c) Asociační tabulka

1. Nechť znaky X_1, …, X_p jsou aspoň ordinálního typu. Sílu pořadové závislosti mezi dvojicemi znaků můžeme posoudit pomocí korelační matice, která obsahuje Spearmanovy koeficienty pořadové korelace. Sílu lineární závislosti mezi dvojicemi znaků můžeme posoudit pomocí korelační matice, která obsahuje výběrové koeficienty korelace.
Význam hodnot korelačního koeficientu:
mezi 0 až 0,1 … zanedbatelná závislost,
mezi 0,1 až 0,3 … slabá závislost,
mezi 0,3 až 0,7 … střední závislost,
mezi 0,7 až 1 … silná závislost.

Statistická indukce: viz přednáška č. 13.

Příklad: Vytvořte korelační matici pro proměnné Výška, Hmotnost, Boty, Pivo, Víno. Na hladině významnosti 0,05 testujte hypotézy o nezávislosti všech dvojic proměnných.

Řešení pomocí systému STATISTICA:

Korelační matice

Proměnná	Vyska	Hmotnost	Boty	Pivo	Vino
Vyska	1,00	0,96	0,96	0,72	-0,14
Hmotnost	0,96	1,00	0,97	0,74	-0,20
Boty	0,96	0,97	1,00	0,70	-0,09
Pivo	0,72	0,74	0,70	1,00	-0,65
Vino	-0,14	-0,20	-0,09	-0,65	1,00

Silný stupeň přímé lineární závislosti existuje mezi proměnnými (Výška, Hmotnost), (Výška, Boty), (Výška, Pivo), (Hmotnost, Boty), (Hmotnost, Pivo), (Boty, Pivo). Střední stupeň nepřímé lineární závislosti existuje mezi proměnnými (Pivo, Víno). Slabý stupeň nepřímé lineární závislosti existuje mezi proměnnými (Výška, Víno), (Hmotnost, Víno). Zanedbatelný stupeň nepřímé lineární závislosti existuje mezi proměnnými (Boty, Víno).
Na hladině významnosti 0,05 se prokázala existence závislosti u dvojic proměnných (Výška, Hmotnost), (Výška, Boty), (Výška, Pivo), (Hmotnost, Boty), (Hmotnost, Pivo), (Boty, Pivo) a (Víno, Pivo).

Řešení pomocí systému SPSS:

Vytvoření korelační matice a tesování hypotézy o nezávislosti všech dvojic proměnných na hladině významnosti 0,05:
Analyze – Correlate – Bivariate – Variables vyska, hmotnost, boty, prijem, pivo, vino – OK.

Correlations

		vyska	hmotnost	boty	prijem	pivo	vino
vyska	Pearson Correlation	1,000	,960**	,961**	,301	,715**	-,138
	Sig. (2-tailed)		,000	,000	,094	,000	,451
	N	32	32	32	32	32	32
hmotnost	Pearson Correlation	,960**	1,000	,969**	,335	,738**	-,197
	Sig. (2-tailed)	,000		,000	,061	,000	,280
	N	32	32	32	32	32	32
boty	Pearson Correlation	,961**	,969**	1,000	,354*	,697**	-,089
	Sig. (2-tailed)	,000	,000		,047	,000	,629
	N	32	32	32	32	32	32
prijem	Pearson Correlation	,301	,335	,354*	1,000	,417*	-,297
	Sig. (2-tailed)	,094	,061	,047		,018	,099
	N	32	32	32	32	32	32
pivo	Pearson Correlation	,715**	,738**	,697**	,417*	1,000	-,654**
	Sig. (2-tailed)	,000	,000	,000	,018		,000
	N	32	32	32	32	32	32
vino	Pearson Correlation	-,138	-,197	-,089	-,297	-,654**	1,000
	Sig. (2-tailed)	,451	,280	,629	,099	,000	
	N	32	32	32	32	32	32

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

2. Vzdálenost mezi objekty můžeme posoudit pomocí matice vzdáleností. Pro znaky intervalového či poměrového typu nejčastěji používáme euklidovskou vzdálenost. Nechť k-tý objekt je popsán vektorem pozorování $\mathbf{x}_k = (x_{k1}, ..., x_{kp})^T$ a l-tý objekt vektorem
$\mathbf{x}_l = (x_{l1}, ..., x_{lp})^T$. Euklidovská vzdálenost k-tého a l-tého objektu:
$d_{kl} = \sqrt{\sum_{j=1}^{p} (x_{kj} - x_{lj})^2}$. Vzdálenosti vypočtené pro všechny dvojice objektů se uspořádají do matice vzdáleností. Je zřejmé, že je to čtvercová symetrická matice, která má na hlavní diagonále nuly.

Příklad: Na pěti objektech byly zjišťovány hodnoty dvou znaků. Datový soubor je tvaru
$\begin{pmatrix} 3 & 7 \\ 5 & 6 \\ 6 & 8 \\ 7 & 10 \\ 9 & 9 \end{pmatrix}$. Najděte matici vzdáleností.

Řešení v systému STATISTICA:
Vytvoříme nový datový soubor o dvou proměnných X_1, X_2 a pěti případech. Zapíšeme do něj zadané hodnoty.

Vytvoření matice euklidovských vzdáleností:
Statistiky – Vícerozměrné průzkumné techniky – Shluková analýza – Spojování (hierarchické shlukování) – OK – Proměnné X1, X2 – OK – na záložce Detaily vybereme Shlukovat Případy (řádky) – OK – na záložce Detaily vybereme Matice vzdáleností.

Případ	P_1	P_2	P_3	P_4	P_5
P_1	0,00	2,24	3,16	5,00	6,32
P_2	2,24	0,00	2,24	4,47	5,00
P_3	3,16	2,24	0,00	2,24	3,16
P_4	5,00	4,47	2,24	0,00	2,24
P_5	6,32	5,00	3,16	2,24	0,00

Vidíme, že nejmenší euklidovskou vzdálenost mají objekty č. 1 a 2, č. 2 a 3, č. 3 a 4, č. 4 a 5.

Řešení v systému SPSS:
Pokud datový soubor vytvořený v systému STATISTICA uložíme s příponou por, můžeme ho otevřít v systému SPSS, jinak obvyklým způsobem vytvoříme nový datový soubor a zapíšeme do něj zadané hodnoty.
Vytvoření matice euklidovských vzdáleností:
Analyze – Classify – Hierarchical Cluster – Variables X1, X2 – Method – Measure Euclidean distance – Continue – OK

Proximity Matrix

Case	Euclidean Distance				
	1	2	3	4	5
1	,000	2,236	3,162	5,000	6,325
2	2,236	,000	2,236	4,472	5,000
3	3,162	2,236	,000	2,236	3,162
4	5,000	4,472	2,236	,000	2,236
5	6,325	5,000	3,162	2,236	,000

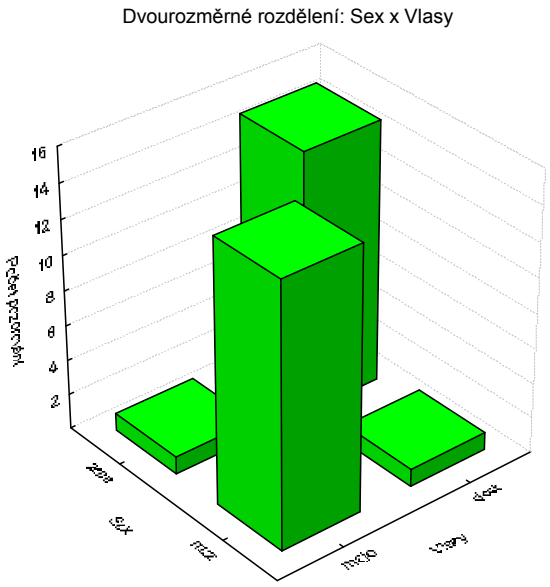
This is a dissimilarity matrix

2. Grafické zpracování

a) 3D sloupkové diagramy
Používají se ke znázornění simultánních absolutních četností v kontingenční tabulce.

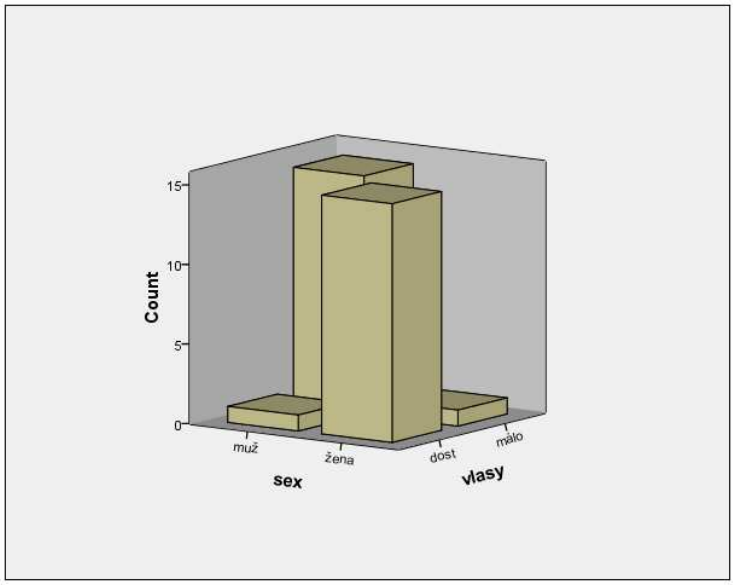
Příklad: Pro proměnné Sex a Vlasy sestrojte 3D sloupkový diagram.

Řešení v systému STATISTICA:



Řešení v systému SPSS:

Graphs – Legacy Dialogs – 3-D Bar – Define – X Category Axis sex,
Z Category Axis vlasy – OK

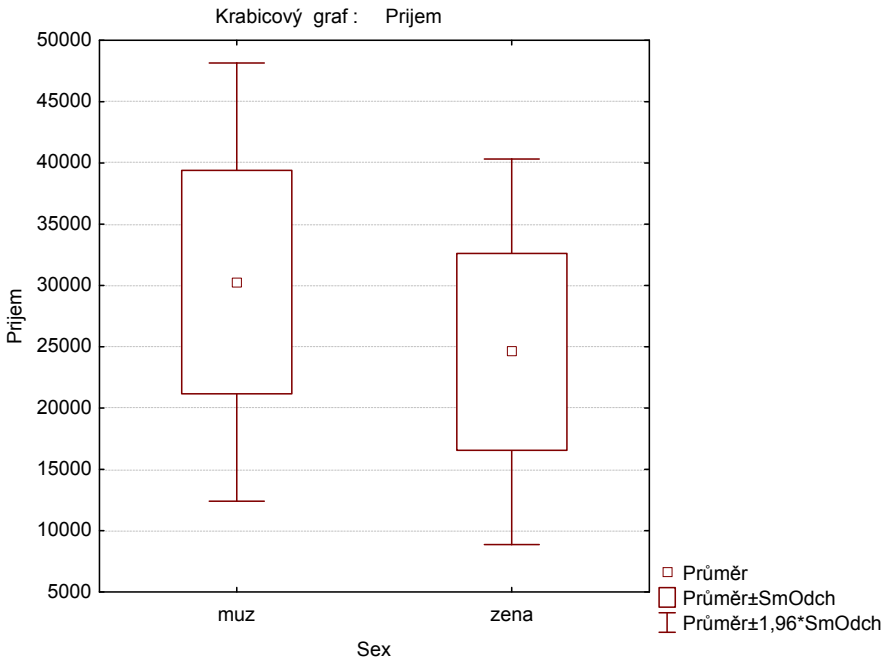


b) Vícenásobné krabicové diagramy

Používají se ke znázornění rozložení dat roztříděných podle úrovní faktoru.

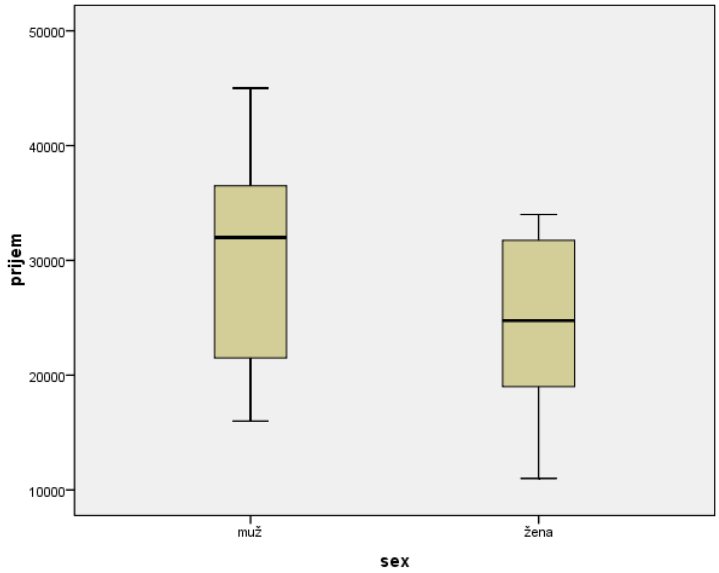
Příklad: Pro proměnnou Příjem roztříděnou podle proměnné Sex sestrojte krabicové diagramy.

Řešení v systému STATISTICA:



Řešení v systému SPSS:

Graph – Legacy Dialogs – Boxplot – Define – Variable prijem, Category Axis sex – OK



Pomocí krabicových diagramů lze snadno detekovat odlehlé či extrémní hodnoty.

Odlehlá hodnota leží mezi vnějšími a vnitřními hradbami, tj. v intervalu $(x_{0,75} + 1{,}5q,\ x_{0,75} + 3q)$ či v intervalu $(x_{0,25} - 3q,\ x_{0,25} - 1{,}5q)$.

Extrémní hodnota leží za vnějšími hradbami, tj. v intervalu $(x_{0,75} + 3q,\ \infty)$ či v intervalu $(-\infty,\ x_{0,25} - 3q)$.

Pomocí nástroje „Průzkumník“ (na liště nástrojů grafu má ikonu lupa ) můžeme v grafu označit názvy objektů, kterým tato odlehlá či extrémní pozorování náleží.

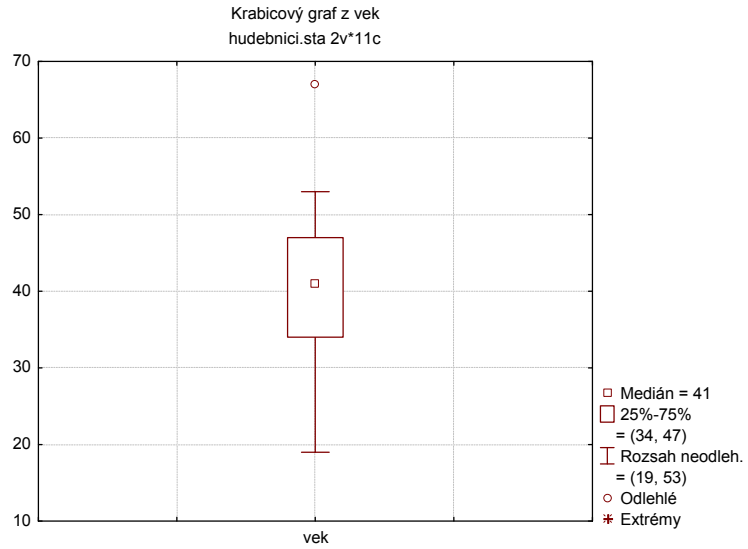
Příklad: Přehlídky dechových hudeb se zúčastnilo 11 hudebníku. Datový soubor obsahuje jejich jména a věk.

1 jmeno	2 vek
Dvořák	53
Šimek	67
Pospíchal	27
Novák	43
Bartoš	19
Kolařík	47
Matoušek	41
Ošmera	34
Němec	34
Fiala	42
Daniel	35

Pomocí krabicového diagramu zjistěte, zda proměnná věk obsahuje odlehlá či extrémní pozorování. Pokud ano, zjistěte jména hudebníků, kterým tato pozorování náleží.

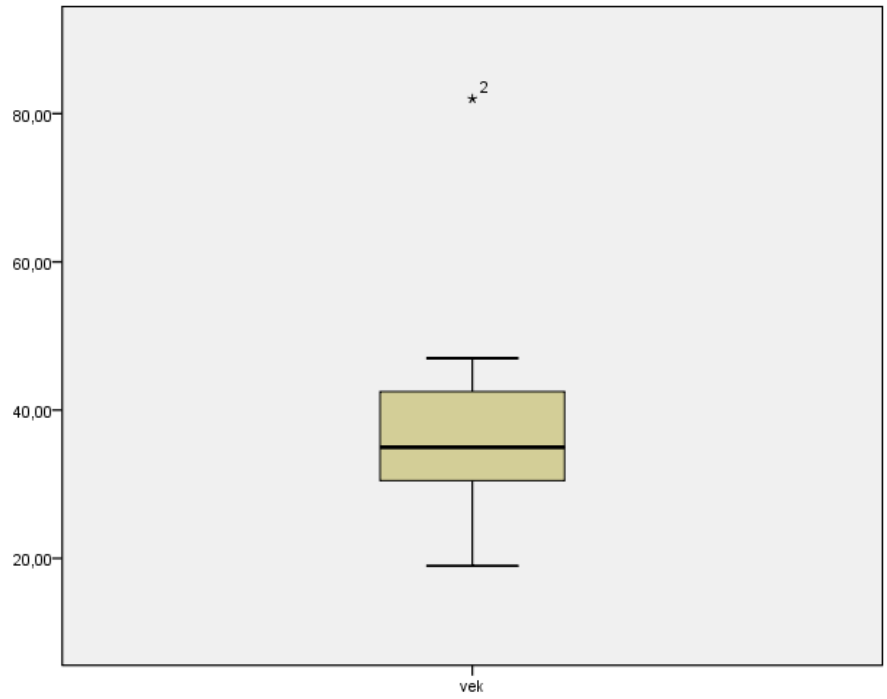
Řešení v systému STATISTICA:

Nejprve označíme případy jmény hudebníků. Data – Správce jmen případů – Přenést jména případů z proměnné jméno – OK – OK. Nyní vytvoříme krabicový diagram pro proměnnou věk: Grafy – 2D Grafy – Krabicové grafy – Proměnné – Závisle proměnné věk – OK – OK.



Vidíme, že v souboru je jedno odlehlé pozorování. Spustíme nástroj Průzkumík (lze tak učinit i z menu: Zobrazit – Průzkumník nebo klikneme pravým tlačítkem na pozadí grafu a vybereme Ukázat průzkumníkem). Zobrazí se lupa a současně se v pravé části obrazovky otevře okno „Průzkumník 2D“. Lupou najedeme na odlehlé pozorování, klikneme na ně myší (tím se pozorování zabarví) a v okně „Průzkumník 2D“ vybereme Použít. U odlehlého pozorování se objeví popis Šimek.

Řešení v systému SPSS:
Graphs – Legacy Dialogs – Boxplot – zaškrtneme Summaries of separate variables – Define – Boxes Represent vek – OK

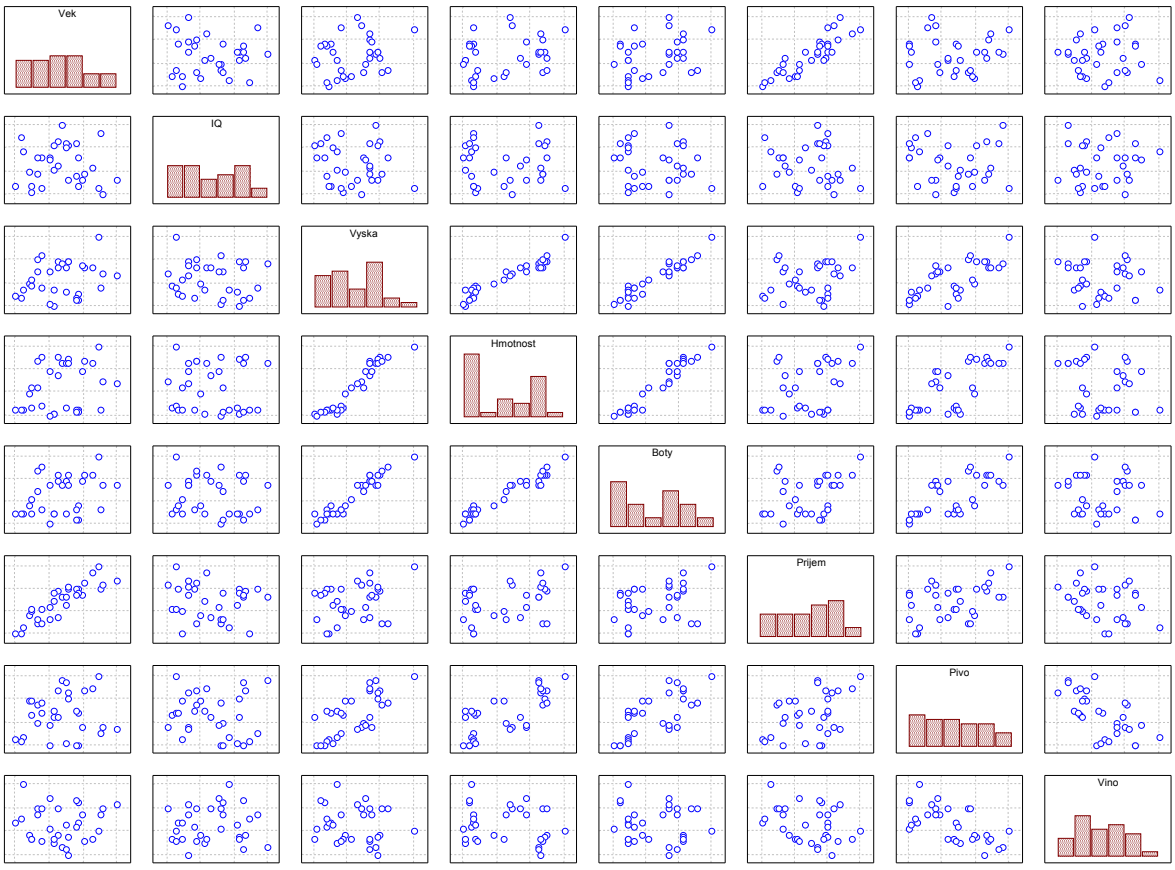


2 x klikneme myší na vyvořený graf. Otevře se Chart Editor. Klikneme pravým tlačítkem na extrémní hodnotu a z menu vybereme Go top Case. Vdíme, že extrémního věku dosahuje hudebník Šimek.

c) Dvourozměrné tečkové diagramy
Používají se ke znázornění závislostí dvojic znaků. Máme-li p znaků, můžeme dvourozměrné tečkové diagramy uspořádat do čtvercového schématu, který se nazývá maticový graf. Na hlavní diagonále jsou histogramy jednotlivých proměnných a mimo hlavní diagonálu jsou dvourozměrné tečkové diagramy příslušných dvojic proměnných.

Příklad: Pro proměnné Věk, IQ, Výška, Hmotnost, Boty, Příjem, Pivo, Víno vytvořte maticový graf.

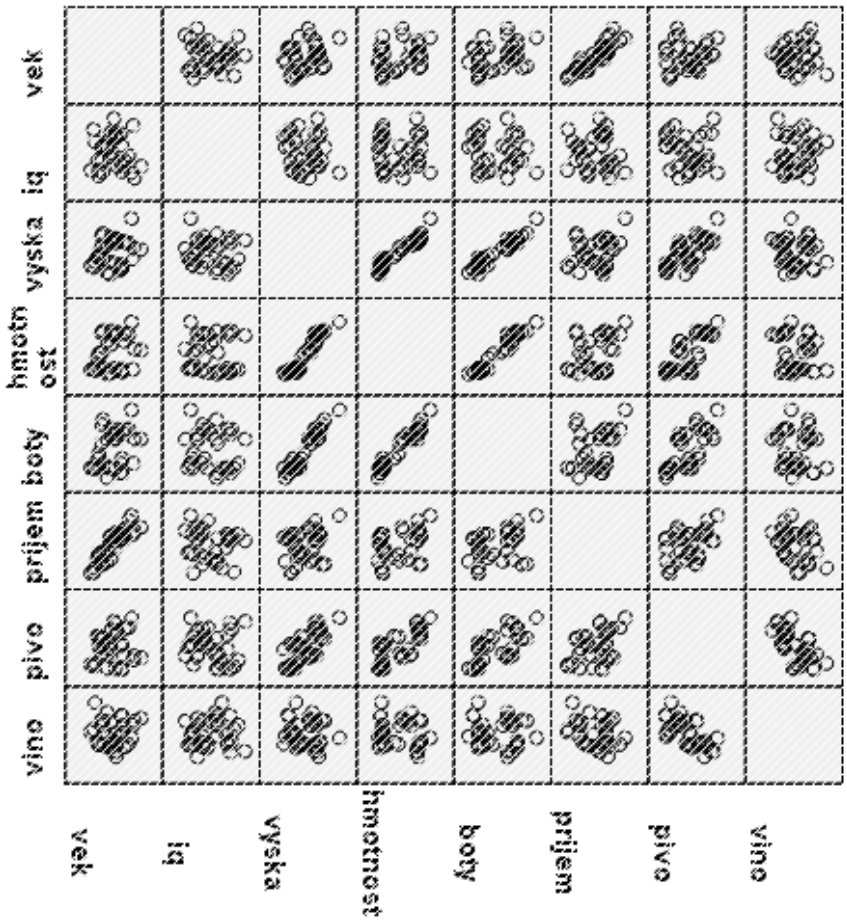
Řešení v systému STATISTICA:
Grafy – Maticové grafy – Proměnné Věk, IQ, Výška, Hmotnost, Boty, Příjem, Pivo, Víno – OK – OK.



Je patrné, že silná přímá lineární závislost existuje mezi proměnnými (Výška, Hmotnost), (Výška, Boty), (Hmotnost, Boty) a (Věk, Příjem). Středně silnou přímou lineární závislost pak vidíme mezi proměnnými (Výška, Pivo), (Hmotnost, Pivo), (Boty, Pivo) a středně silnou nepřímou lineární závislost pak mají proměnné (Pivo, Víno).

Řešení v systému SPSS

Vytvoření maticového grafu:
Graphs – Legacy Dialogs – Scatter/dot – Matrix Scatter – Define – Matrix Variables vek, iq, vyska, hmotnost, boty, prijem, pivo, vino – OK



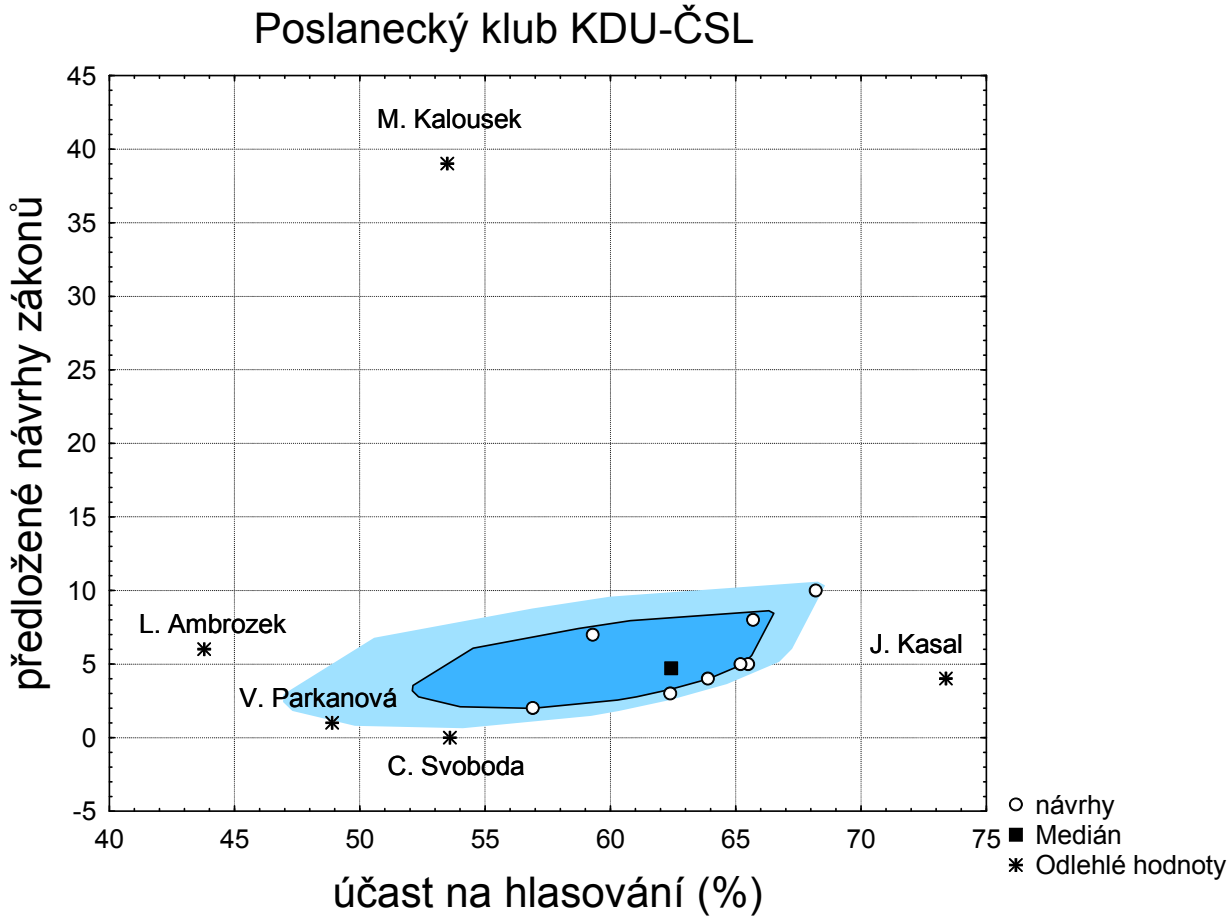
d) Bag plot

Jedná se o typ dvourozměrného tečkového diagramu užívající zobecnění krabicového grafu k identifikaci rozložení a odlehlých hodnot v dvourozměrném prostoru. Jeho aplikaci si ukážeme na datech z Poslanecké sněmovny Parlamentu ČR.

Na stránce www.psp.cz jsou dostupné údaje o jednotlivých poslancích, např. o počtu návrhů zákonů, které poslanec podal a o jeho účasti na hlasování. (Data pocházejí z 15.10.2008, tedy zachycují stav do 38. schůze PSP ČR včetně.) Podíváme se na vztah mezi těmito dvěma veličinami u poslanců KDU – ČSL.

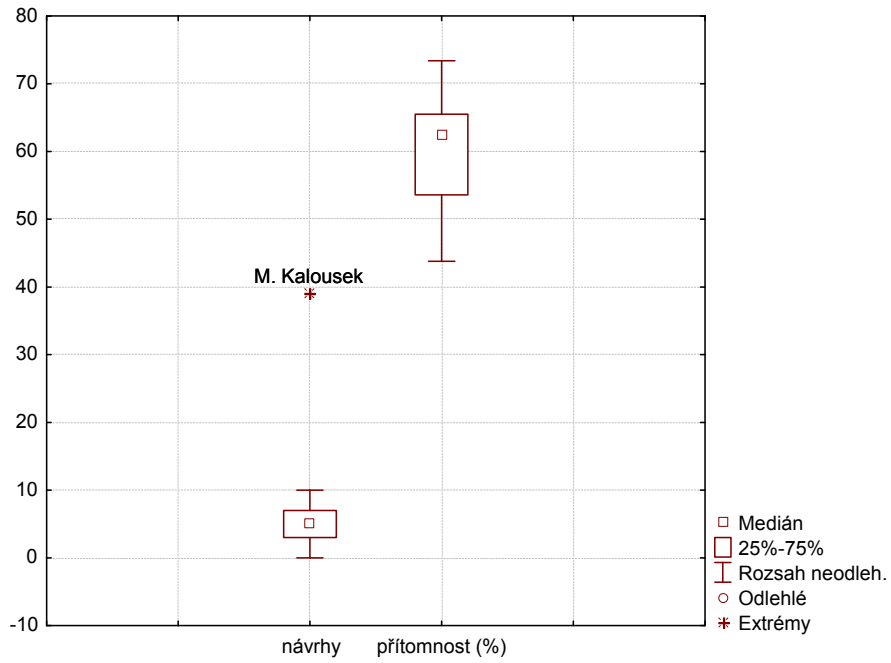
	1 návrhy	2 přítomnost (%)
L. Ambrozek	6	43,8
J. Carbol	5	65,5
J. Hanuš	10	68,2
L. Hovorka	5	65,2
M. Kalousek	39	53,5
J. Kasal	4	73,4
T. Kvapil	8	65,7
V. Parkanová	1	48,9
P. Severa	2	56,9
C. Svoboda	0	53,6
M. Šimonovský	4	63,9
M. Šojdrová	7	59,3
L. Šustr	3	62,4

Řešení v systému STATISTICA:



Poslanci jsou v grafu označeni kolečkem. Odlehlé hodnoty v dvourozměrném prostoru jsou označeny hvězdičkou. Tmavě modrá oblast (bag) odpovídá krabici klasického krabicového grafu s mediánem a kvartily. Uvnitř této oblasti leží 50% pozorování. Světle modrá oblast reprezentuje svorky klasického krabicového grafu, uvnitř kterých leží neodlehlé hodnoty.
Z grafu je okamžitě vidět, kteří poslanci KDU – ČSL se ocitají mimo „hlavní proud“. Např. Miroslav Kalousek předložil 39 návrhů zákonů, ale jeho účast na hlasování byla jen 53,5%. Naproti tomu Jan Kasal měl účast ze všech poslanců KDU – ČSL nejvyšší (73,4%), předložil však jenom 4 návrhy zákonů. Vlasta Parkanová předložila 1 návrh a měla účast 48,9%, což je druhá nejnižší po Liborovi Ambrozkovi (43,8%, 6 návrhů zákonů). Cyril Svoboda nepředložil žádný návrh zákona a jeho účast na hlasování činila 53,6%.

Pro srovnání se podíváme na aktivitu poslanců pomocí obyčejného krabicového diagramu.



V počtu předložených návrhů zákonů byl nejaktivnější Miroslav Kalousek, jehož 39 předložených návrhů představuje dokonce extrémní hodnotu. Co se týká účasti na hlasování, zde se nevyskytují žádné odlehlé ani extrémní hodnoty. Je tedy zřejmé, že bag plot umožňuje komplexnější pohled na dvourozměrná data než obyčejný krabicový diagram.

e) Ikonové (symbolové) grafy

Hodnoty znaků jsou převedeny do určitých geometrických útvarů nebo symbolů. Každému objektu pak odpovídá jistý obrazec složený z těchto útvarů či symbolů. Vyhodnocení dat pak provedeme srovnáním těchto obrazců, např. hledáním podobných obrazců. K nejpoužívanějším symbolovým grafům patří profilové sloupce, profily a Chernoffovy tváře.

Profilové sloupce: Ke každému objektu je sestrojena soustava sloupců, jejichž výšky odpovídají relativním hodnotám uvažovaných znaků (relativní hodnota vznikne jako podíl původní hodnoty a maxima z absolutních hodnot znaku).

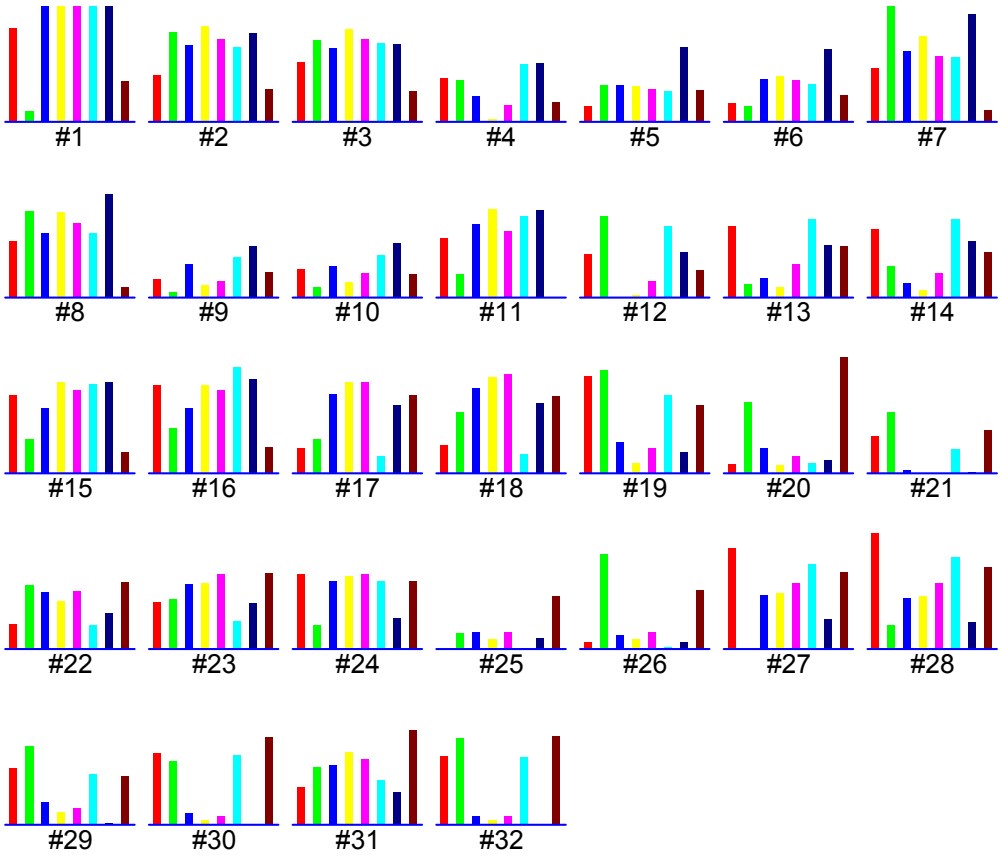
Profily: Středy horních hran profilových sloupců se spojí úsečkami.

Chernoffovy tváře: charakterizují každý znak nějakým prvkem schématizovaného obličeje, např. šířkou obličeje, délkou nosu, šířkou úst, zakřivením úst apod. Vzhled tváře samozřejmě závisí na použitém pořadí znaků.

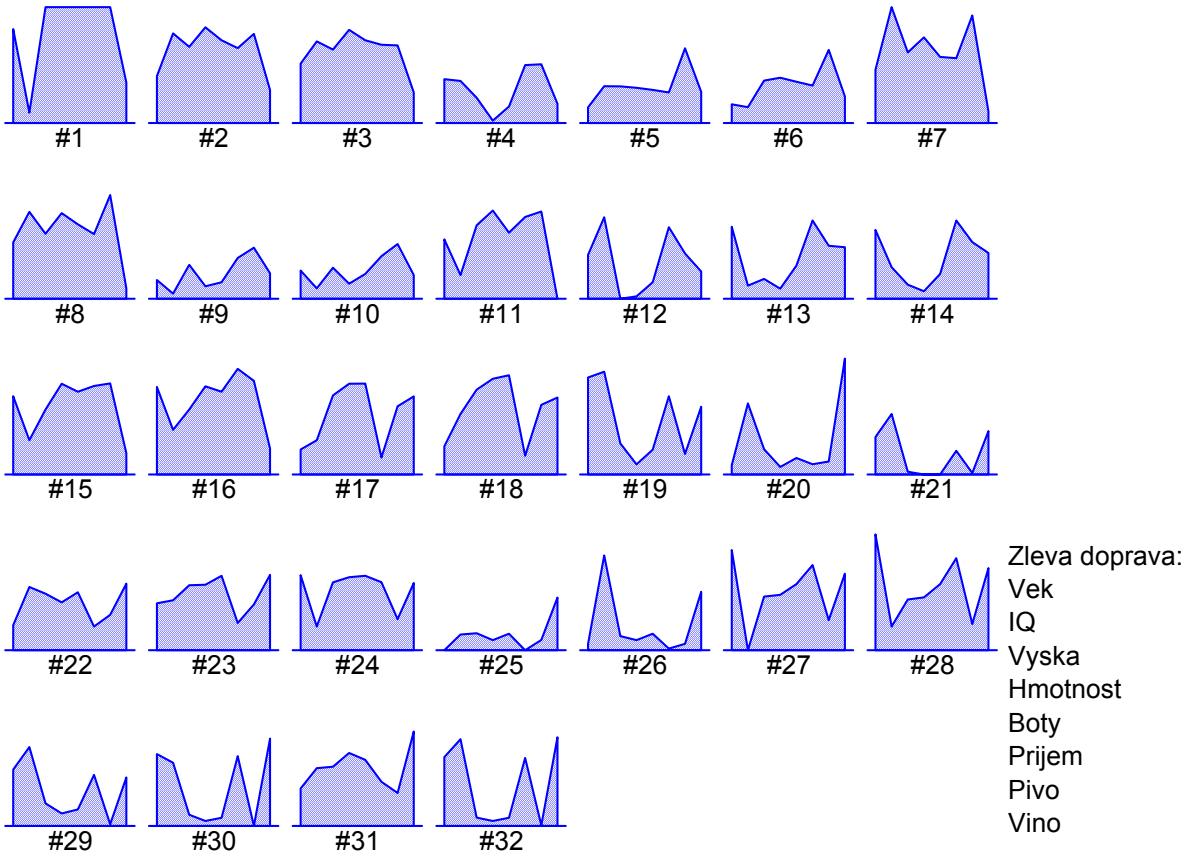
Příklad: Vytvořte sloupce, profily a Chernoffovy tváře pro proměnné Věk, IQ, Výška, Hmotnost, Bota, Příjem, Pivo, Víno z datového souboru Lidé.

Řešení v systému STATISTICA:

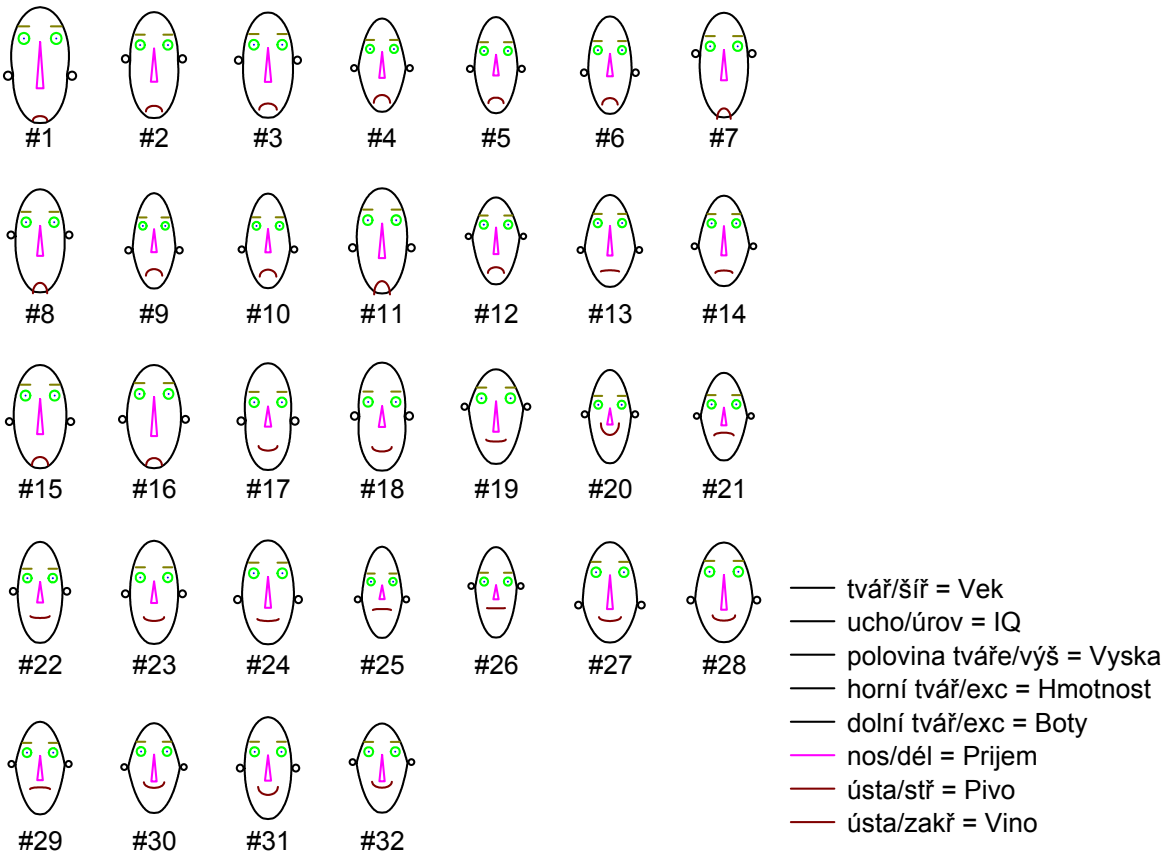
Profilové sloupce: Grafy – Ikonové grafy – Proměnné Věk, IQ, Výška, Hmotnost, Bota, Příjem, Pivo, Víno – OK, Typ grafu Sloupce – Možnosti 1 – zapnout Zobrazit popisy případů, zvolit Jména případů



Profily: V Typu grafu zvolíme Profily



Chernoffovy tváře: V Typu grafu zvolíme Chernoffovy tváře


#1
#2
#3
#4
#5
#6
#7
#8
#9
#10
#11
#12
#13
#14
#15
#16
#17
#18
#19
#20
#21
#22
#23
#24
#25
#26
#27
#28
#29
#30
#31
#32
tvář/šíř = Vek
ucho/úrov = IQ
polovina tváře/výš = Vyska
horní tvář/exc = Hmotnost
dolní tvář/exc = Boty
nos/dél = Prijem
ústa/stř = Pivo
ústa/zakř = Vino