Kontingenční tabulky

(Analýza kategoriálních dat)

Agenda

Standardní analýzy dat v kontingenčních tabulkách

- úvod, KT, míry diverzity nominálních veličin, některá rozdělení
- chí kvadrát testy, analýza reziduí, Fisherův přesný test, 4 polní tabulka, míry asociace
- metody zobrazení dat z kont. tabulek a závislostí
- loglineární modely

Další techniky - asociační pravidla, stromy a grafické modely

Základní pojmy

- Kategoriální data; nominální, ordinální
- Reprezentace v kontingenční tabulce
- Pro dvě proměnné dostaneme např. 2x3 tabulku:

	Genotyp			
	AA	**Aa**	**aa**	
Nemoc +	n_{11}	n_{12}	n_{13}	$n_{1.}$
Nemoc -	n_{21}	n_{22}	n_{23}	$n_{2.}$
	$n_{.1}$	$n_{.2}$	$n_{.3}$	n

$$n=n_{11}+n_{12}+n_{13}+n_{21}+n_{22}+n_{23}$$

Míry diverzity nominální veličiny

- vazba na míry závislosti a vytváření klasifikačních modelů
- Zkoumaný znak nabývá hodnot $A_1, \ldots, A_k$ s pravděpodobnostmi $p_1, \ldots, p_k$
- *(Shannonova)* ***entropie*** $$H = -\sum_{j=1}^{k} p_j \log_2(p_j)$$
- ***Giniho index*** $$1 - \sum_{j=1}^{k} p_j^2 = \sum_{j=1}^{k} p_j (1 - p_j)$$

Příklad:

relativní četnosti kategorií jsou 0.5 a 0.5

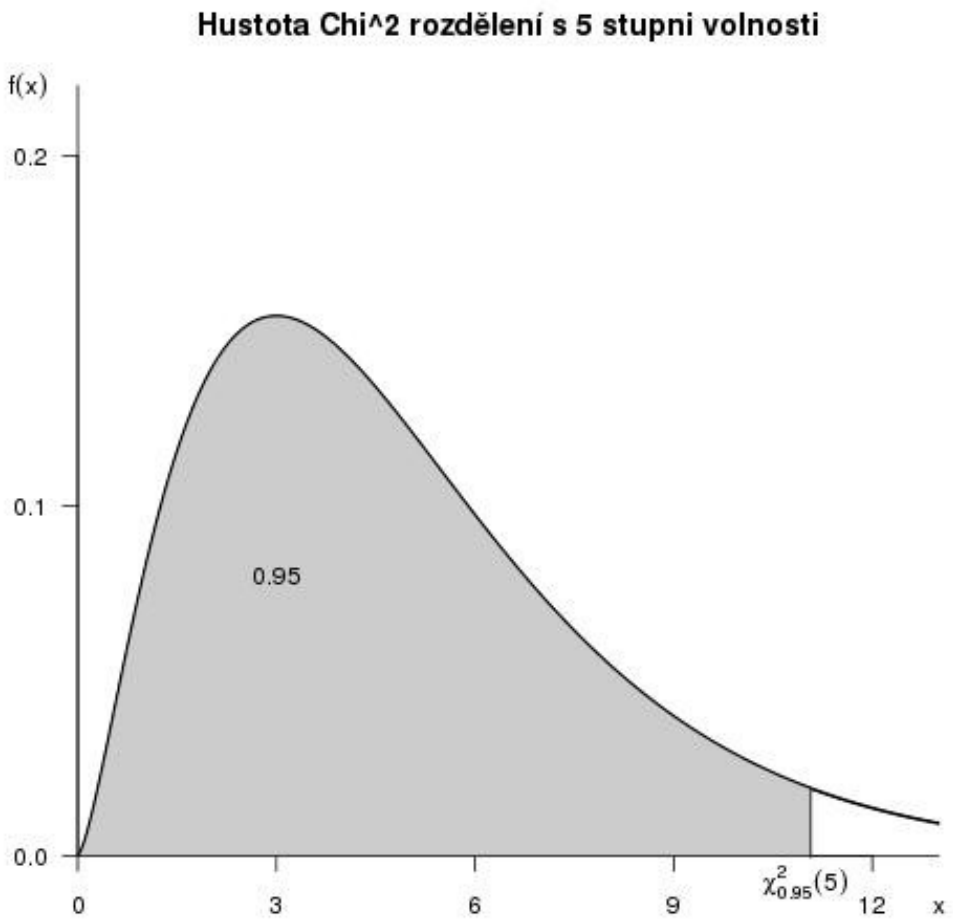
H= ... $$H = -(0.5\log_2(0.5) + 0.5\log_2(0.5)) = -(-0.5 + (-0.5)) = 1$$

zákl. pojmy – multinomické rozdělení

- n nezávislých pokusů, v každém jsou možné výsledky $A_1, \ldots, A_k$ (disjunktní a vyčerpávající) s pravděpodobnostmi $\pi_1, \ldots, \pi_k$ (pro $k=2$ jde o binomické rozdělení)
- Četnosti $N_1, \ldots, N_k$ těchto výsledků mají multinomické rozdělení
- Pro každou k-tici nezáporných čísel $n_1, \ldots, n_k$, kde $\sum_{j=1}^{k} n_j = n$

$$P(N_1 = n_1, \ldots, N_k = n_k) = \frac{n!}{n_1! \ldots n_k!} \pi_1^{n_1} \ldots \pi_k^{n_k}$$

Rozdělení χ^2



χ^2 test dobré shody

- hypotéza určuje všechny pravděpodobnosti:
- $N_1, \ldots, N_k$ je náhodný vektor s multinomickým rozdělením s parametry $n, \pi_1, \ldots, \pi_k$,

$$H_0 : \pi_1 = \pi_1^0, \ldots, \pi_k = \pi_k^0$$

- Spočteme teoretické četnosti $o_i = n*\pi_i$
- Porovnáme teoretické a skutečné četnosti $(o_i \geq 5)$
- Testová statistika $$X^2 = \sum_{i=1}^{k} \frac{(N_i - o_i)^2}{o_i}$$
- H_0 zamítneme, je-li $$X^2 \geqslant \chi^2_{1-\alpha}(k-1)$$

χ^2 test nezávislosti

- H_0 *:* nezávislost dvou nominálních veličin A,B nebo shoda pravděpodobností v několika populacích
- n_{ij} četnost dvojice hodnot: i-té hodnoty *A* a zároveň j-té hodnoty *B*; marginální četnosti $n_{i.}$,$n_{.j}$
- teoretické četnosti (za předpokladu nezávislosti) $o_{ij} = n_{i.}\, n_{.j} / n$
- Porovnáme teoretické a skutečné četnosti $(o_i \geq 5)$
- Testová statistika $$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{\left(N_{ij} - o_{ij}\right)^2}{o_{ij}}$$
- H_0 zamítneme, je-li $$X^2 \geqslant \chi^2_{1-\alpha}\left(\left(r-1\right)*\left(c-1\right)\right)$$

Příklad – test nezávislosti

uvažujme 2x3 tabulku

	AA	Aa	aa
Znak B: ano	6	14	20
Znak B: ne	13	9	33

n=6+14+20+13+9+33=95

(6+14+20)(6+13)/95=8

	AA	Aa	aa
Znak B: ano	8	10	22
Znak B: ne	11	13	31

- tabulka očekávaných četností
- statistika $X^2=\frac{(6-8)^2}{8}+\frac{(14-10)^2}{10}+\frac{(20-22)^2}{22}+\frac{(13-11)^2}{11}+\frac{(9-13)^2}{13}+\frac{(33-31)^2}{31}$
- hodnotu $X^2 = 4$ porovnáváme s $\chi^2_{(0.95)}(df=(3-1)(2-1)=2)=6$

a hypotézu o nezávislosti nezamítáme

Čtyřpolní tabulka

Pro dvě dvouhodnotové proměnné dostaneme 2x2 tabulku

	Znak A var 1	**Znak A not var 1**
Znak B var 1	a	b
Znak B not var 1	c	d

$n=a+b+c+d$

Klasická statistika

$$X^2=\frac{n(ad-bc)^2}{(a+b)(a+c)(b+d)(c+d)}$$

Yatesova korekce

$$X_Y^2=\frac{n(|ad-bc|-n/2)^2}{(a+b)(a+c)(b+d)(c+d)}$$

Malé počty pozorování

(výpočetně náročnější) řešení problému nízkých teoretických četností (a neplatnosti odvozování podle klasické statistiky)

- **Fisherův** (přesný) test $p_a = \frac{(a+b)!(a+c)!(b+d)!(c+d)!}{n!a!b!c!d!}$

p_a je pravděpodobnost konkrétní tabulky (2x2) při daných marginálních četnostech

sečteme pravděpodobnost dané tabulky a tabulek ještě více odporujících nulové hypotéze a dostaneme p hodnotu testu

- **simulace** s využitím generátoru pseudonáhodných čísel

Míry asociace nominálních veličin

Hledáme obdobu korelačního koeficientu,

vypovídající o těsnosti/síle závislosti

Pro 4polní tabulku:

- **poměr šancí** (šance jako P(A)/(1-P(A))) $\mathrm{OR} = \frac{ad}{bc}$

$$S.E.\left(\ln\left(\mathrm{OR}\right)\right) = \sqrt{1/a + 1/b + 1/c + 1/d}$$

přibližný interval spolehlivosti pro logaritmus populačního podílu šancí

(ln(OR) − S.E.(ln(OR))z(α/2), ln(OR) + S.E.(ln(OR))z(α/2))

Míry asociace nominálních veličin / 2

- pro 4-polní tabulku leží mezi 0 a 1 koeficient $\phi = \sqrt{X^2/n}$

- Cramerovo V $V = \sqrt{\frac{X^2}{n(m-1)}}$

kde $m = min(r,c)$

Analýza reziduí

Rezidua: $r_{ij} = n_{ij} - o_{ij}$

Standardizovaná rezidua: $sr_{ij} = \frac{r_{ij}}{\sqrt{o_{ij}}}$

Adjustovaná std. rezidua: $asr_{ij} = \frac{r_{ij}}{\sqrt{o_{ij}\left(1-\frac{n_{i.}}{n}\right)\left(1-\frac{n_{.j}}{n}\right)}}$

$n = a+b+c+d$

$n_{1.} = a+b$

$n_{.2} = b+d$

Očekávané Četnosti o_{ij}	**Znak A – var 1**	**Znak A- not var 1**
Znak B – var 1	$n_{1.} * n_{.1} / n$	$n_{1.} * n_{.2} / n$
Znak B – not var 1	$n_{2.} * n_{.1} / n$	$n_{2.} * n_{.2} / n$

Vizualizace závislostí u kategoriálních dat

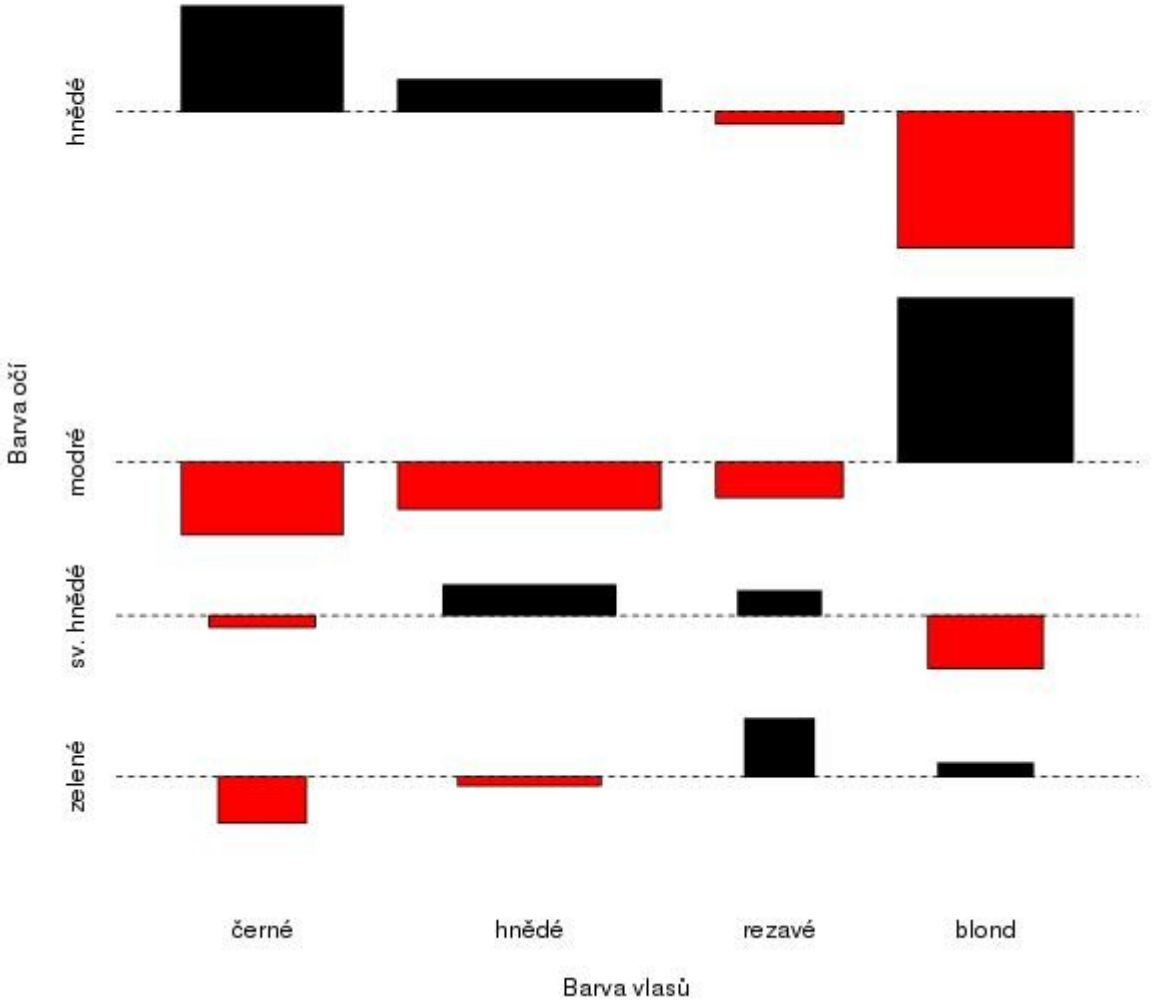
Zobrazení závislostí na úrovni proměnných → grafické modely,...
Zobrazení závislostí „vnitřní struktury“ tabulky (na úrovni kategorií)

- Mosaic plot
- Association plot

- Meyer, D., Zeileis, A., and Hornik, K. (2005) The strucplot framework: Visualizing multi-way contingency tables with vcd. Report 22, Department of Statistics and Mathematics, Wirtschaftsuniversität Wien, Research Report Series. http://epub.wu-wien.ac.at/dyn/openURL?id=oai:epub.wu-wien.ac.at:epub-wu-01_8a1

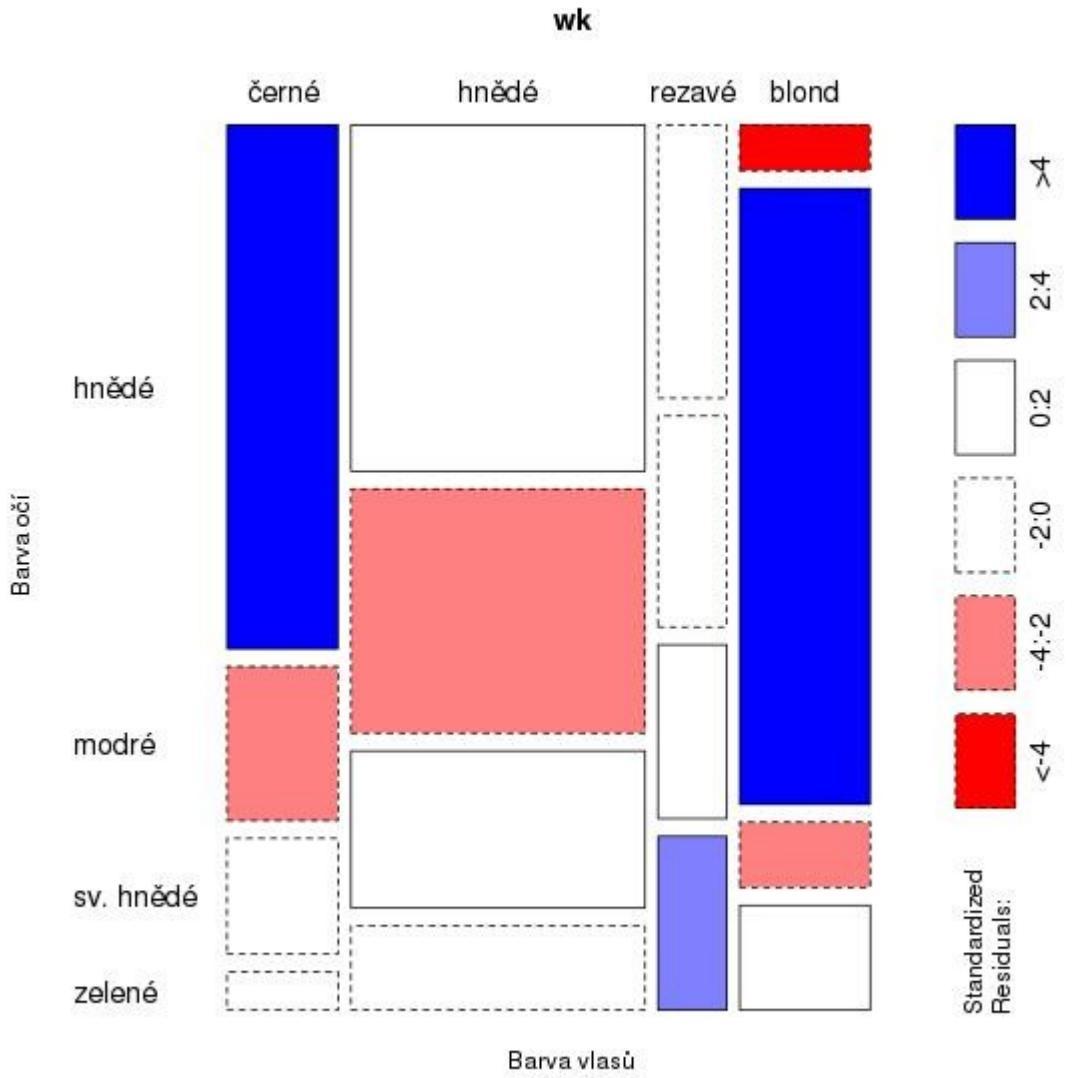
Vizualizace - „graf asociací“

Association plot



Vizualizace - „mozaikový graf“

mosaic plot



Loglineární modely

Modelují četnosti v kontingenční tabulce

- Pro dvě proměnné A (řádek), B (sloupec):
- Model nezávislosti $o_{ij} = n p_{i.} p_{.j}$

$$\log o_{ij} = \log n + \log p_{i.} + \log p_{.j}$$

$$\lambda_i^A = \log p_{i.} - \left(\sum_h \log p_{h.}\right)/I \qquad \mu = \log n + \left(\sum_h \log p_{h.}\right)/I + \left(\sum_h \log p_{.h}\right)/J$$

$$\lambda_j^B = \log p_{.j} - \left(\sum_h \log p_{.h}\right)/J \qquad \sum \lambda_i^A = \sum \lambda_j^B = 0$$

$$\log o_{ij} = \mu + \lambda_i^A + \lambda_j^B$$

- Saturovaný model $\log o_{ij} = \mu + \lambda_i^A + \lambda_j^B + \lambda_{ij}^{AB}$

Explorační analýza dat a data mining

Analýza rozsáhlých dat v situacích, kdy není moc jasné, co může být výsledkem

Nevíme přesně na co se ptát:

„Jsou v datech nějaké **zajímavé** vztahy?“

(x konfirmační analýza dat, ve které ověřujeme hypotézu)

Asociační pravidla

Automaticky (počítačem) generovat všechny hypotézy zajímavé na základě daných empirických dat

- Sledujeme více kategoriálních proměnných současně
- Vznik kolem aplikací zaměřených na analýzu nákupního košíku (dichotomické proměnné)
- Snaha objevit často se vyskytující kombinace znaků „frequent itemsets“
- Výpočetně náročné postupy
- Možnost zadat obecnou podobu vztahu, který nás zajímá
- Možnost zadat požadavky na minimální spolehlivost, podporu a podobně

Někdy obtížné vyhodnocení výsledků

Asociační pravidla / 2

Různé logické tvary hypotézy, „φ souvisí s ψ“, kde φ a ψ jsou kombinace atributů

- Například „Jestliže – pak“ konstrukce:
- Antecedent -> sukcedent
- Závěr (sukcedent) není předem určen
- Počet zkoumaných kombinací při neomezené analýze *m* proměnných je $\prod_{j=1}^{m}\left(1+K_{A_j}\right)-1$

Jaký je vztah mezi spolehlivostí pravidla
A & B & C - > D a pravidla A & B -> C & D ??

Asociační pravidla - charakteristiky kvality

- ***Podpora*** *(support) =P(Ant & Suc) = a/(a+b+c+d)*
  a je podíl případů splňujících předpoklad i závěr pravidla, někdy se uvádí také absolutní podpora (*a*)

- ***Spolehlivost*** *P(Suc|Ant)= a/(a+b)*
  podmíněná pravděpodobnost závěru, platí-li předpoklad

- ***Pokrytí*** P(Ant | Suc)=a / (a+c)

- ***Kvalita*** = w_1**spolehlivost* + w_2**pokrytí*

- **Konzistentní** pravidla – spolehlivost = 1, Ant je PP závěru
- **Úplná** pravidla – pokrytí = 1, Ant je nutná podmínka závěru
  deterministické pravidlo = konzistentní a úplné

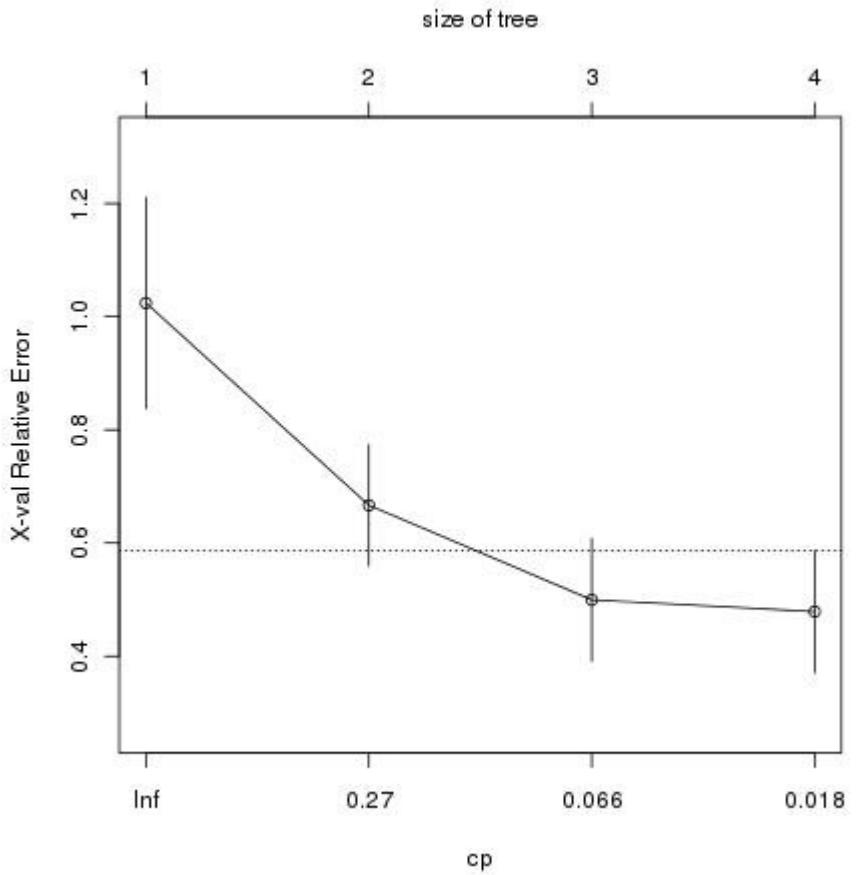
Klasifikační: stromy

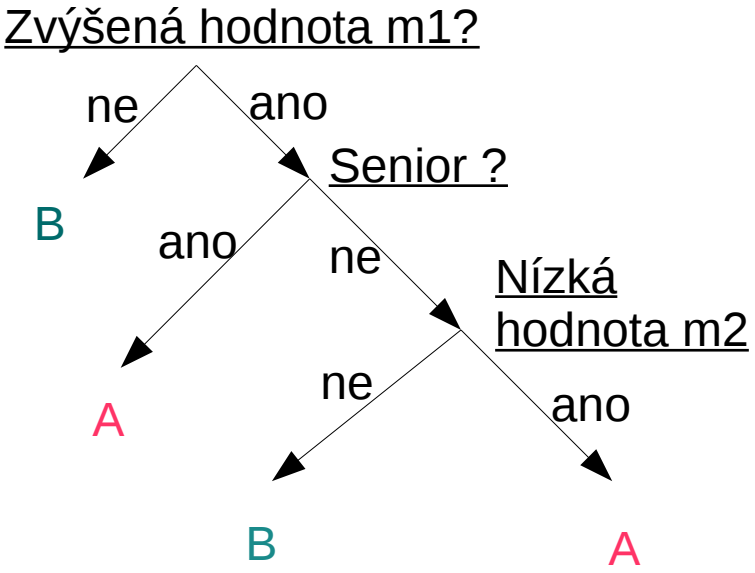
- Cílem je klasifikace případu podle atributů
- Vytváření stromu: rekurzivní rozklad vstupních dat podle nejlépe rozlišující proměnné

Výhody metody:

- Možnost zachytit složitější interakce, vztahy platné jen pro určitou podskupinu
- Prakticky žádné předpoklady o datech; pro kategoriální i spojitá data, chybějící hodnoty
- Výsledek modelování je (někdy) přehledný, snadná interpretace
- Použitelné pro identifikaci důležitých proměnných

Ukázka analytických technik - rozhodovací strom





Grafický model - bayesovská síť

Orientovaný acyklický graf (uzel odpovídá náhodné veličině) a sada pravděpodobnostních fcí – pro každý uzel *U* ve tvaru *P(U|rodiče(U))*

Faktorizace sdružené pravděpodobnostní funkce (řetězové pravidlo):

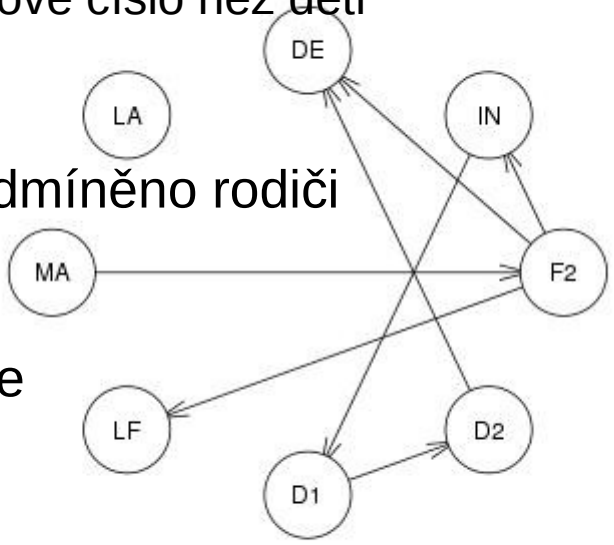
P(LA,MA,LF,D1,D2,IN,DE, F2) =

P(LA)P(MA)P(F2|MA)P(LF|F2)P(IN|F2)P(D1|IN)P(D2|D1)P(DE|D2,F2)

V grafu očíslujeme všechny veličiny tak, aby rodiče měli nižší pořadové číslo než děti

Pak každá veličina je podmíněně nezávislá na všech

veličinách s nižším pořadovým číslem mimo svých rodičů podmíněno rodiči



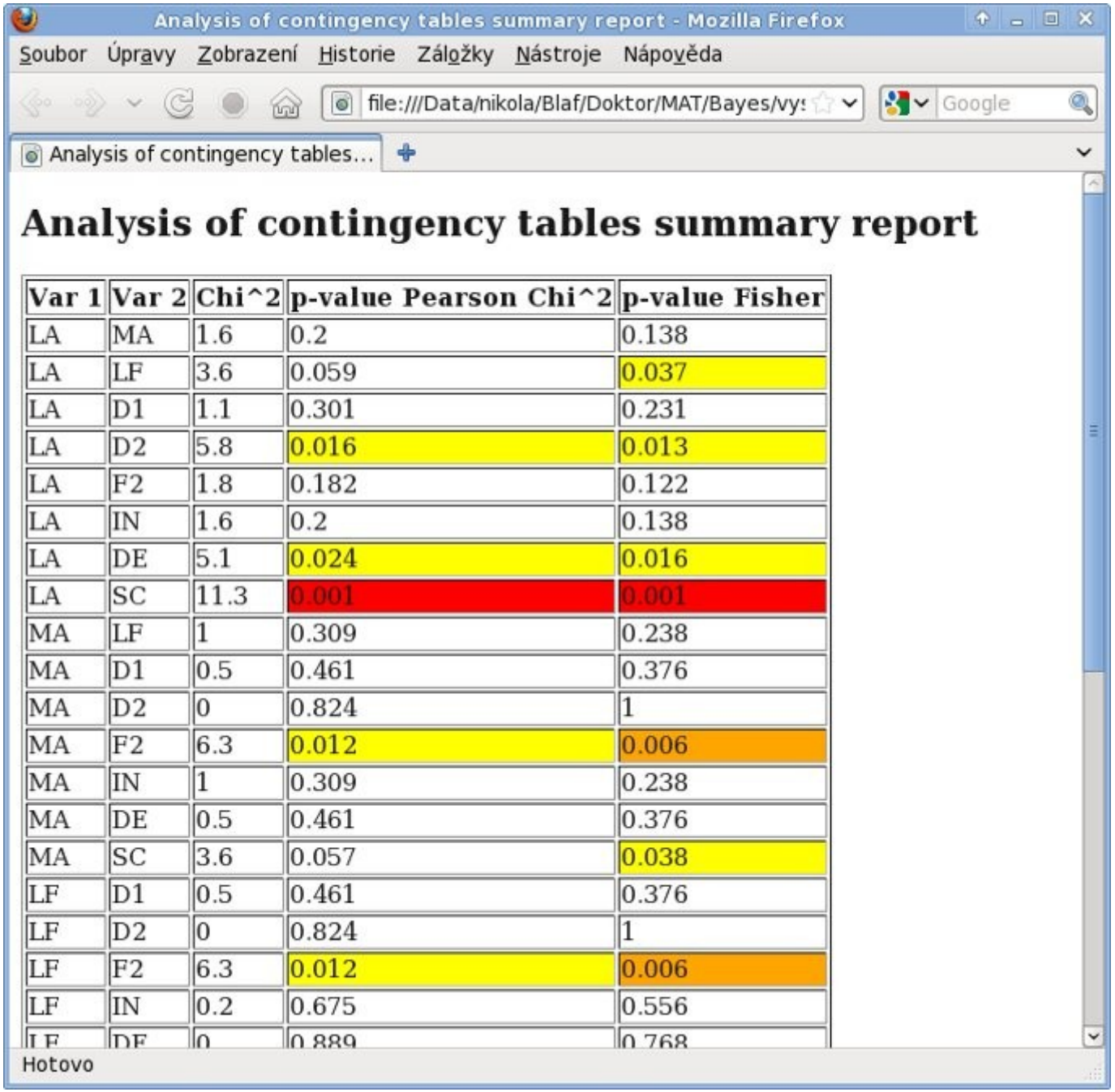
Veličiny A a B jsou podmíněně nezávislé při daném C, jestliže

P(A,B|C)=P(A|C)*P(B|C)

Ekvivalentní vztah

P(A|B,C)=P(A|C), P(B|A,C)=P(B|C)

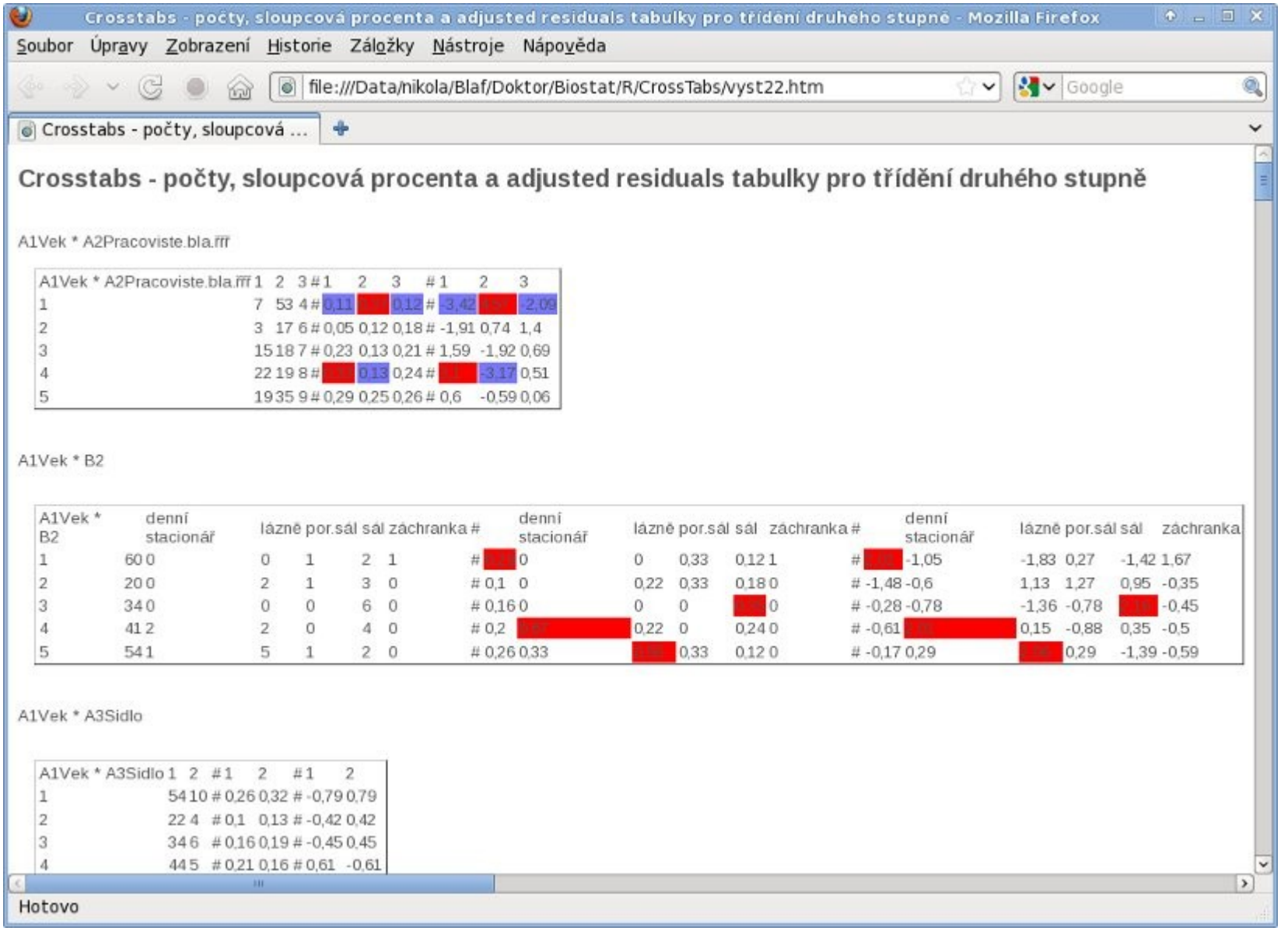
Report z analýzy kont. tabulek



Analysis of contingency tables summary report

Var 1	Var 2	Chi^2	p-value Pearson Chi^2	p-value Fisher
LA	MA	1.6	0.2	0.138
LA	LF	3.6	0.059	0.037
LA	D1	1.1	0.301	0.231
LA	D2	5.8	0.016	0.013
LA	F2	1.8	0.182	0.122
LA	IN	1.6	0.2	0.138
LA	DE	5.1	0.024	0.016
LA	SC	11.3	0.001	0.001
MA	LF	1	0.309	0.238
MA	D1	0.5	0.461	0.376
MA	D2	0	0.824	1
MA	F2	6.3	0.012	0.006
MA	IN	1	0.309	0.238
MA	DE	0.5	0.461	0.376
MA	SC	3.6	0.057	0.038
LF	D1	0.5	0.461	0.376
LF	D2	0	0.824	1
LF	F2	6.3	0.012	0.006
LF	IN	0.2	0.675	0.556
LF	DE	0	0.889	0.768

Report z analýzy kont. tabulek - 2



Crosstabs - počty, sloupcová procenta a adjusted residuals tabulky pro třídění druhého stupně - Mozilla Firefox

Soubor Úpravy Zobrazení Historie Záložky Nástroje Nápověda

file:///Data/nikola/Blaf/Doktor/Biostat/R/CrossTabs/vyst22.htm

Google

Crosstabs - počty, sloupcová ...

Crosstabs - počty, sloupcová procenta a adjusted residuals tabulky pro třídění druhého stupně

A1Vek * A2Pracoviste.bla.fff

A1Vek * A2Pracoviste.bla.fff	1	2	3	#	1	2	3	#	1	2	3
1	7	53	4	#	0,11	0,5	0,12	#	-3,42	3,57	-2,09
2	3	17	6	#	0,05	0,12	0,18	#	-1,91	0,74	1,4
3	15	18	7	#	0,23	0,13	0,21	#	1,59	-1,92	0,69
4	22	19	8	#	0,3	0,13	0,24	#	2,1	-3,17	0,51
5	19	35	9	#	0,29	0,25	0,26	#	0,6	-0,59	0,06

A1Vek * B2

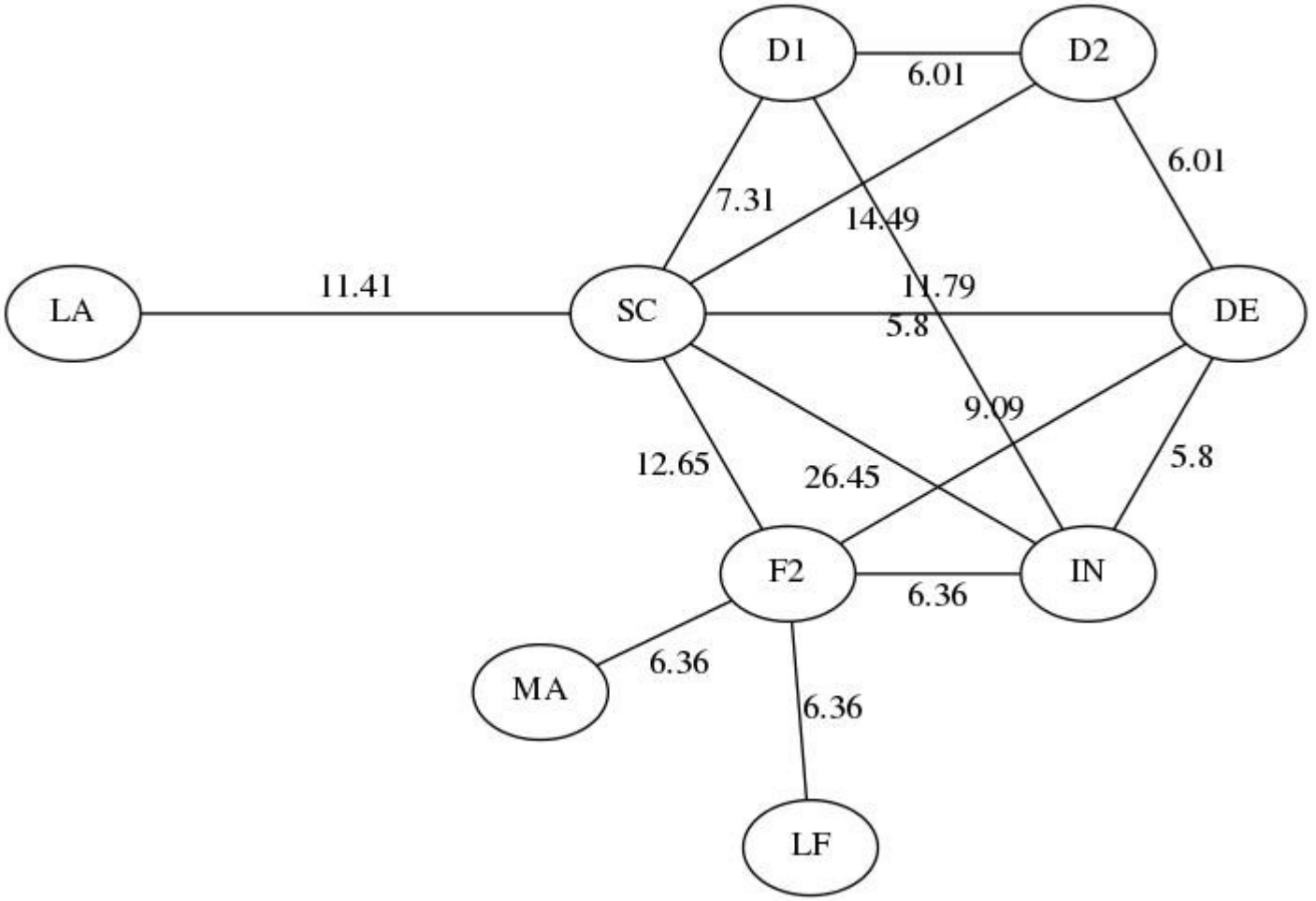
A1Vek * B2	denní stacionář	lázně	por.sál	sál	záchranka	#	denní stacionář	lázně	por.sál	sál	záchranka	#	denní stacionář	lázně	por.sál	sál	záchranka
1	6	0	0	1	2	1	#	1,1	0	0	0,33	0,12	1	#	0,45	-1,05	-1,83	0,27	-1,42	1,67
2	2	0	0	2	1	3	0	#	0,1	0	0,22	0,33	0,18	0	#	-1,48	-0,6	1,13	1,27	0,95	-0,35
3	3	4	0	0	0	6	0	#	0,16	0	0	0	0,35	0	#	-0,28	-0,78	-1,36	-0,78	1,16	-0,45
4	4	1	2	2	0	4	0	#	0,2	0,67	0,22	0	0,24	0	#	-0,61	1,61	0,15	-0,88	0,35	-0,5
5	5	4	1	5	1	2	0	#	0,26	0,33	0,44	0,33	0,12	0	#	-0,17	0,29	-0,46	0,29	-1,39	-0,59

A1Vek * A3Sidlo

A1Vek * A3Sidlo	1	2	#	1	2	#	1	2
1	54	10	#	0,26	0,32	#	-0,79	0,79
2	22	4	#	0,1	0,13	#	-0,42	0,42
3	34	6	#	0,16	0,19	#	-0,45	0,45
4	44	5	#	0,21	0,16	#	0,61	-0,61

Hotovo

Report pro 2x2 tabulky
- graf závislostí podle Fisherova testu, OR



...

mail: data@tulipany.cz
Web: http://skola.tulipany.cz