
Statistické výpočetní prostředí R

A některé možnosti aplikací pro analýzu biomedicínských dat

Agenda

Úvodní představení R – co je R a odkud se vzalo

- ukázka: možnost tvorby GUI (Tcl/Tk)
- grafika v R
- automatizace reportů a grafů, generování zdroje pro Graphviz

Nástroje pro explorační analýzu dat a data mining

Balíčky a Bioconductor – nástroje pro bioinformatiku

Představení R – historie

R je open source implementací programovacího **jazyka S**, další známou implementací je komerční software S-PLUS

- ACM Software System Award 1998
John M. Chambers (Bell Labs) „For The S system, which has forever altered how people analyze, visualize, and manipulate data“

ACM SSA - „Awarded to an institution or individual(s) recognized for developing a software system that has had a lasting influence, reflected in contributions to concepts, in commercial acceptance, or both“

odbočka: Za co se dá dostat ACM ocenění

2009 – VMware Workstation for Linux 1.0	1994 – Remote Procedure Call
2008 – The Gamma Parallel Database System	1993 – Sketchpad
2007 – Statemate	1992 – Interlisp
2006 – Eiffel	1991 – TCP/IP
2005 – The Boyer-Moore Theorem Prover	1990 – NLS
2004 – Secure Network Programming	1989 – PostScript
2003 – MAKE	1988 – INGRES
2002 – Java	1988 – System R
2001 – SPIN	1987 – SMALLTALK
1999 – Apache	1986 – TeX
1997 – Tcl/Tk	1985 – VisiCalc
1995 – NCSA Mosaic	1984 – Xerox Alto System
1995 – World-Wide Web	1983 – UNIX

Představení R

Prostředí pro programování, analýzu dat, statistické výpočty a grafiku, dobře pokrývá všechny standardní typy analýz

- Open source software, licence GNU GPL
- Vznik kolem r. 1995, velká rozšířenost, integrace v komerčních statistických softwarech
- Navrhován jako efektivní a spolehlivý (dokumentace i vývoj)
- Funguje na různých platformách – OS Linux / Unix a další
- Flexibilní, snadno rozšiřitelný nástroj
- Efektivní balíčkovací systém, množství knihoven
- Široká komunita uživatelů a otevřený kód dávají dobré předpoklady pro rozvoj systému

web: <http://www.r-project.org>

Představení R / 2 -tech.

- plnohodnotný interpretovaný programovací jazyk
- možnost využívat principů objektově orientovaného programování (generické funkce, třídy, metody, dědičnost), podpora abstraktních datových typů, výstupem výpočtu je objekt použitelný pro další programování
- (téměř) funkcionální programovací jazyk
- možnost využít kód v jiných jazycích (C, Fortran, ...), interface do databázových systémů
- podpora konceptu reproducible research (systém Sweave)
 - integruje programový kód pro zpracování dat a dokument
 - standardně příkazová řádka, podpora dávkového zpracování, existují GUI nadstavby, podpora Tcl/Tk (ukázka)

Představení R / 3 - tech. - další rozvoj

Vlastnosti jazyka R mohou být pro některé účely omezující – proto existují vývojové aktivity, které se tato omezení snaží obejít

- využití striktně funkcionálního jazyka, jazyka s dostupnými kvalitními překladači, typovaného jazyka
- jako perspektivní se jeví tradiční Lisp

Představení R / 4 - balíčky, Bioconductor

- Množství dostupných balíčků (knihoven)
- Sada nástrojů Bioconductor – (od r. 2002) projekt vyvíjený s podporou Fred Hutchinson Cancer Research Center
“an open source and open development software project for the analysis and comprehension of genomic data.”

Podpora úloh moderní biostatistiky a bioinformatiky, včetně náročné fáze přípravy dat, analýza microarrays experimentů, ...

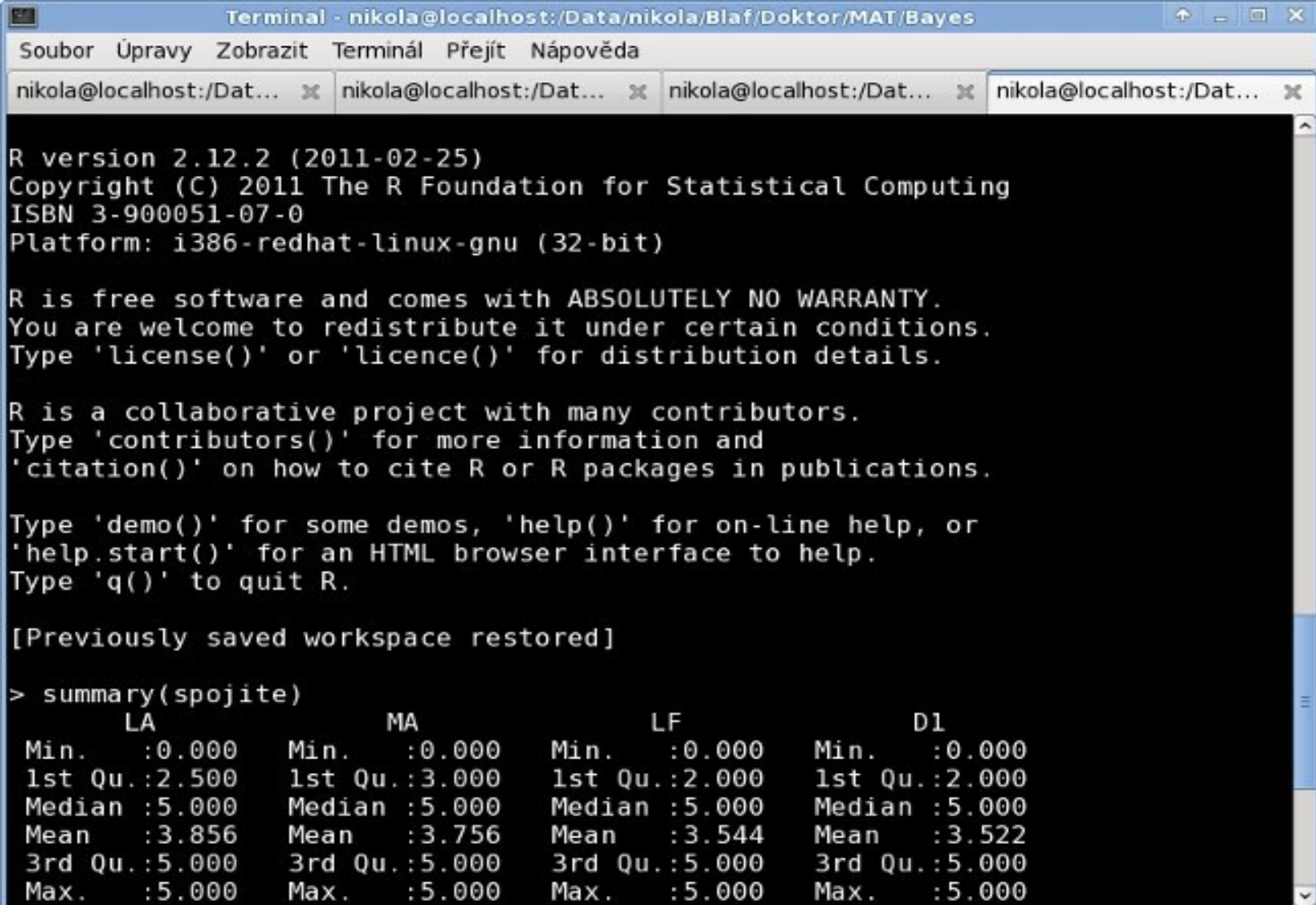
Datové struktury schopné pojmout i metadata – sebe-popisující objekty

Podpora různých datových formátů, včetně např. FASTA formátu pro data o biologických sekvencích

Propojení na online biomedicínské zdroje – NCBI, PubMed, GO, ...

Jak může R vypadat?

... příkazová řádka



```
Terminal - nikola@localhost:/Data/nikola/Blaf/Doktor/MAT/Bayes
Soubor Úpravy Zobrazit Terminál Přejít Nápověda
nikola@localhost:/Dat... x nikola@localhost:/Dat... x nikola@localhost:/Dat... x nikola@localhost:/Dat... x

R version 2.12.2 (2011-02-25)
Copyright (C) 2011 The R Foundation for Statistical Computing
ISBN 3-900051-07-0
Platform: i386-redhat-linux-gnu (32-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[Previously saved workspace restored]

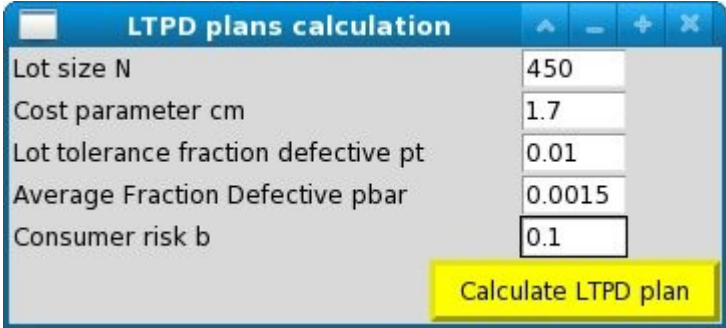
> summary(spojite)
      LA           MA           LF           D1
Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   :0.000
1st Qu.:2.500   1st Qu.:3.000   1st Qu.:2.000   1st Qu.:2.000
Median :5.000   Median :5.000   Median :5.000   Median :5.000
Mean   :3.856   Mean   :3.756   Mean   :3.544   Mean   :3.522
3rd Qu.:5.000   3rd Qu.:5.000   3rd Qu.:5.000   3rd Qu.:5.000
Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000
```

Jak může R vypadat / 2

... GUI vyvinuté na míru

Analogie příkazu

funkce(Argument1, Argument2, Argument3, Argument4, Argument5)



The screenshot shows a window titled "LTPD plans calculation" with a blue header bar. It contains five input fields with the following values: Lot size N (450), Cost parameter cm (1.7), Lot tolerance fraction defective pt (0.01), Average Fraction Defective pbar (0.0015), and Consumer risk b (0.1). A yellow button labeled "Calculate LTPD plan" is located at the bottom right of the window.

Lot size N	450
Cost parameter cm	1.7
Lot tolerance fraction defective pt	0.01
Average Fraction Defective pbar	0.0015
Consumer risk b	0.1

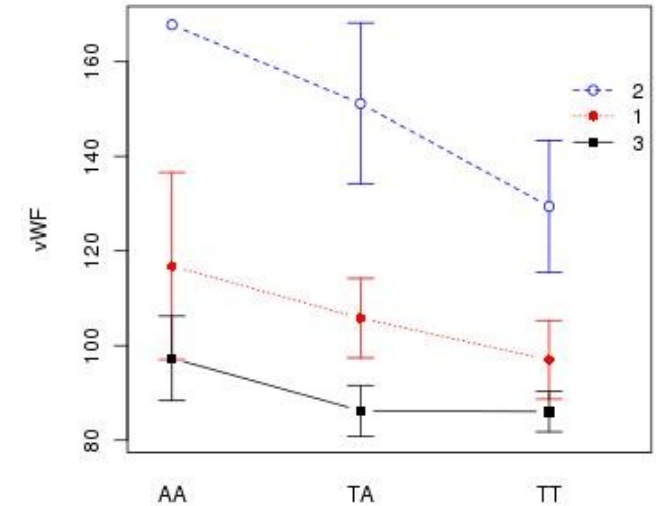
Grafické uživatelské rozhraní může uživateli odlehčit od nutnosti orientovat se v syntaxi ... ale ztratíme tím některé výhody

Grafika v R

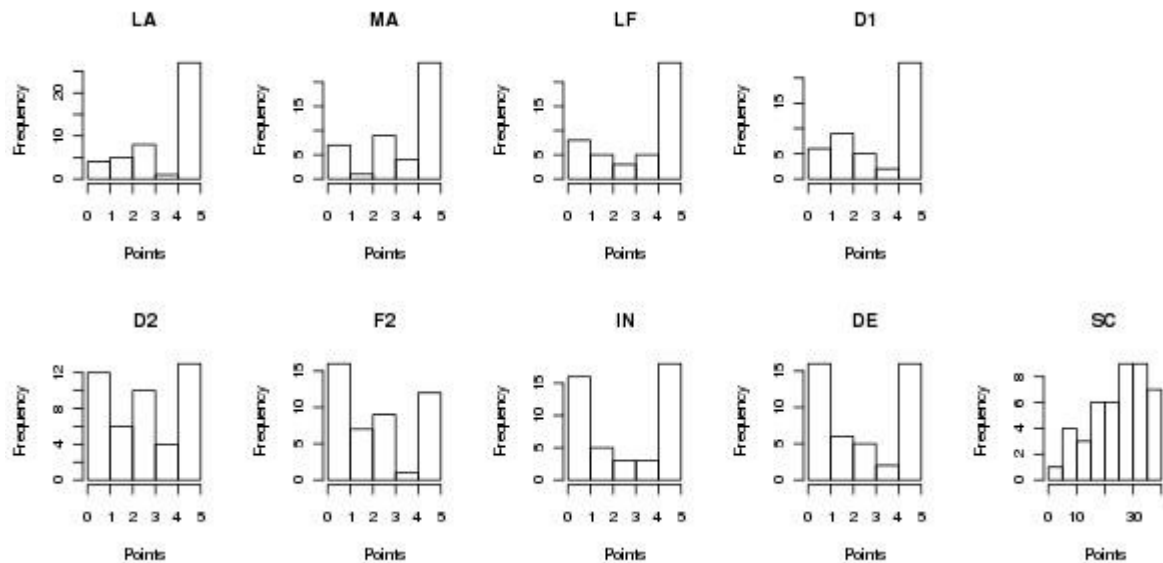
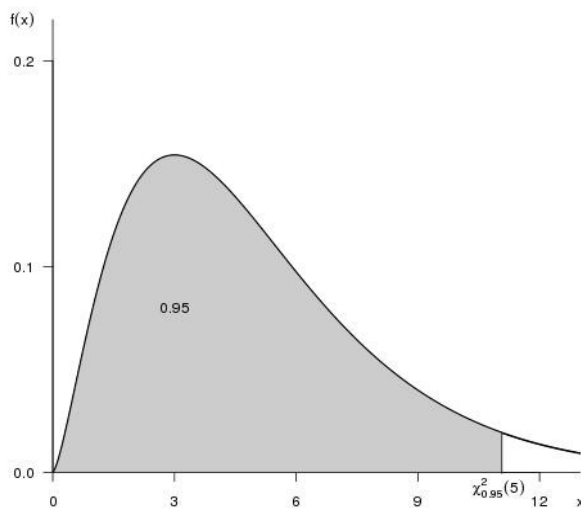
Tradiční systém, grid a lattice

Velice flexibilní

Možnost výstupu v různých formátech (.jpg, xfig, .emf, PostScript, .pdf, ...)



Hustota χ^2 rozdělení s 5 stupni volnosti



Explorační analýza dat a data mining

e. a. d. – volné vymezení konceptu volné analýzy:

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

Nevíme přesně na co se ptát**:

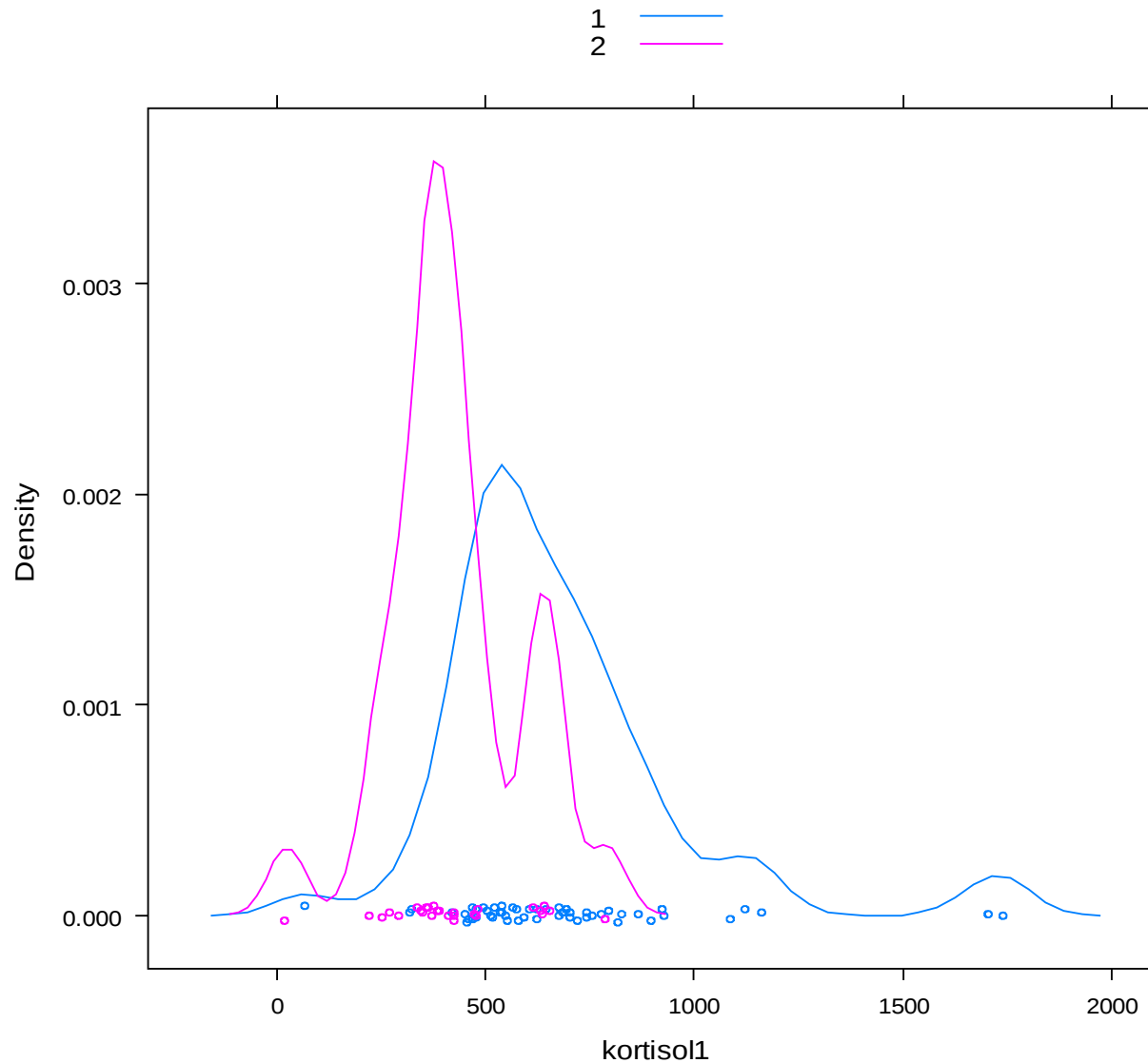
„Jsou v datech nějaké **zajímavé** vztahy?“

(x konfirmační analýza dat, ve které ověřujeme hypotézu)

*Někdy není moc jasné také jak naložit s výsledkem analýzy

**You get no answer if you have no question

Ukázka analytických technik - graf hustoty pravděpodobnosti



Explorační analýza dat: stromy

Pro klasifikaci (klasifikační stromy) a predikci spojitých proměnných (regresní stromy)

Rekurzivní rozklad vstupních dat podle nejlépe rozlišující proměnné

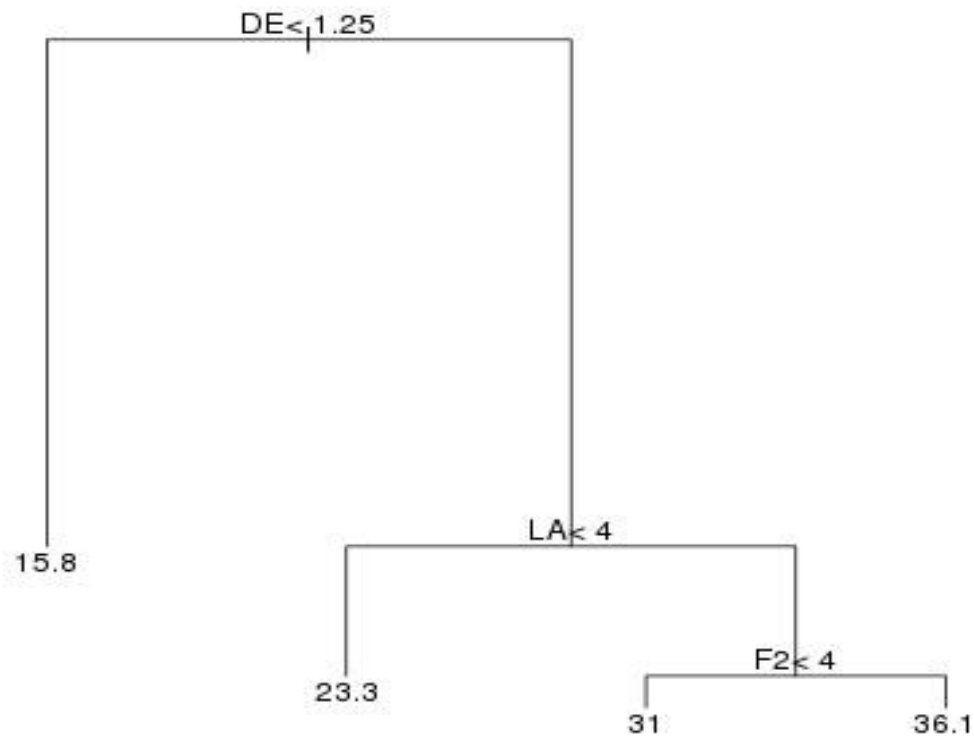
Možnost zachytit složitější interakce, vztahy platné jen pro určitou podskupinu

Prakticky žádné předpoklady o datech; pro kategoriální i spojitá data, chybějící hodnoty

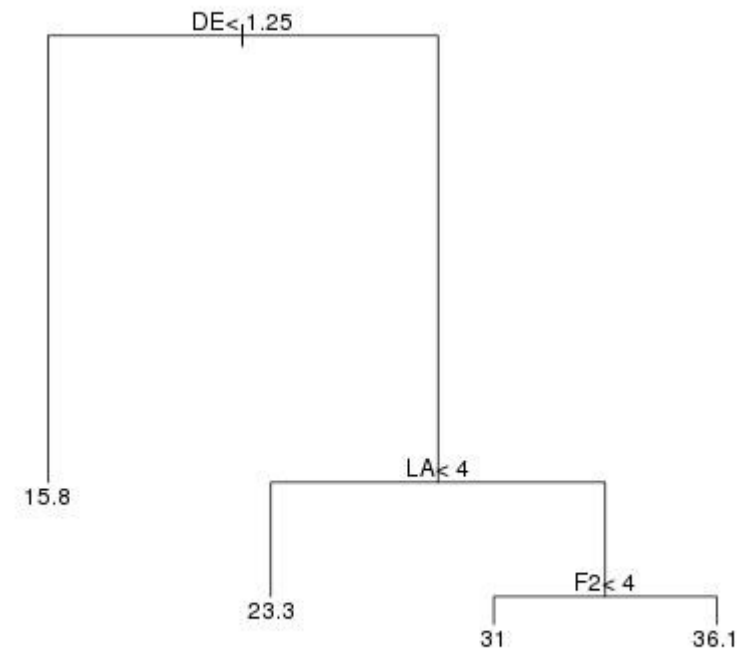
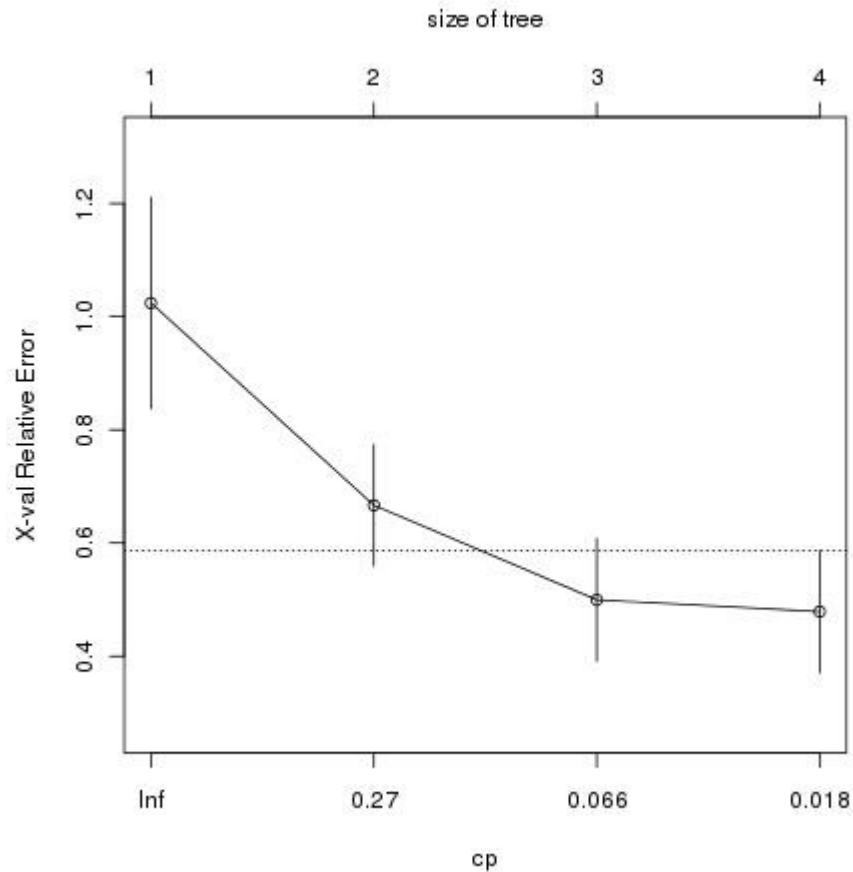
Výsledek modelování je (někdy) přehledný, snadná interpretace

Použitelné pro identifikaci důležitých proměnných

Ukázka analytických technik - regresní strom



Ukázka analytických technik - regresní strom - vyhodnocení



Ukázka analytických technik - bayesovská síť

Bayesian network learned via Score-based methods

model:

[LA][MA][F2|MA][LF|F2][IN|F2][D1|IN][D2|D1][DE|D2:F2]

nodes: 8

arcs: 7

undirected arcs: 0

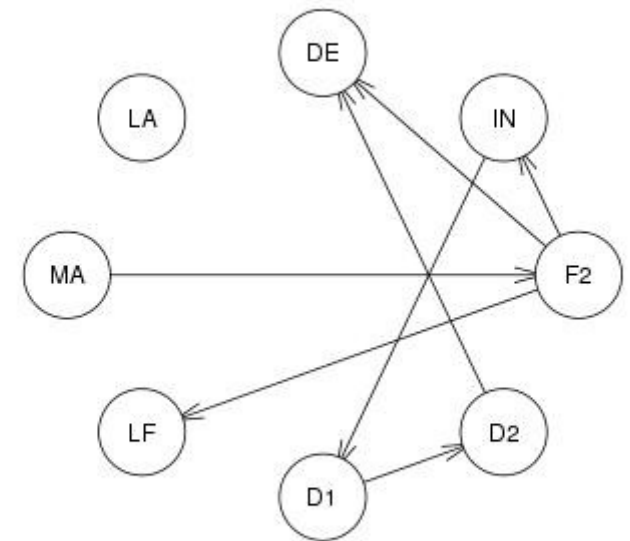
directed arcs: 7

learning algorithm: Hill-Climbing

score: Bayesian Information Criterion

penalization coefficient: 1.903331

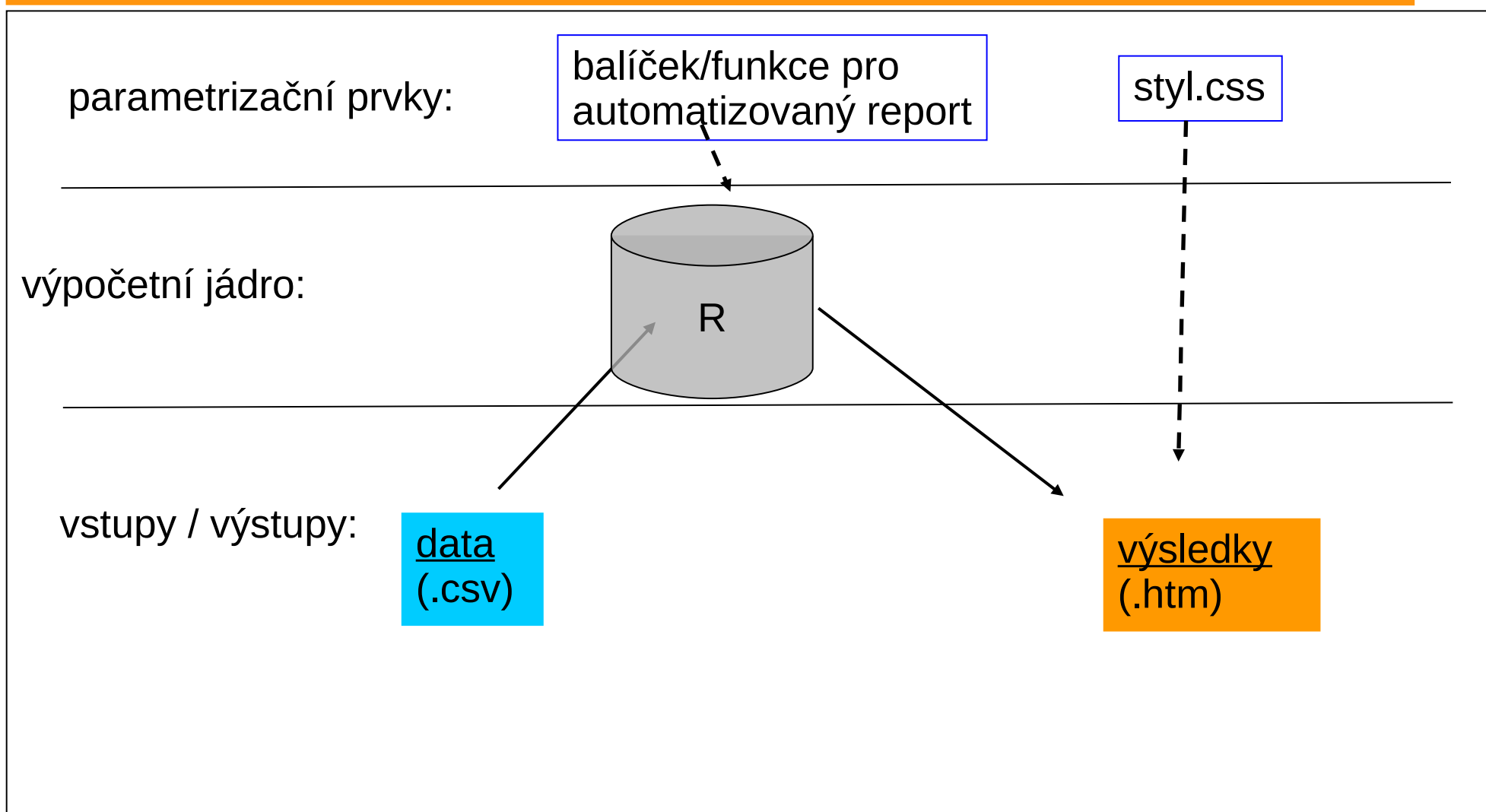
tests used in the learning procedure: 392



Automatizovaný reporting

- arch. návrh založen na sledovaných principech
 - modularita – řešení je sadou jednotlivých komponent, které jsou poměrně nezávislé (-> přehledné a snadno přístupné úpravám), s jasně definovaným vstupem a výstupem
 - řešení postavené na otevřených standardech, nezávislé na platformě: využití formátu CSV (soubor s daty oddělenými středníkem) pro vstup a značkovacího jazyka HTML pro výstup = formáty, u kterých nedochází k nepříjemným překvapením
 - větší parametrizace?
- data k analýze mám v tabulkovém procesoru ...
 - > v tabulkovém procesoru soubor uložit ve formátu CSV – data oddělená středníkem, desetinný oddělovač čárka
- co když chci výsledky mít v excelu??
 - > otevřít si .htm výstup -> označit vše (Ctrl-A) -> kopírovat (Ctrl-C) -> vložit (Ctrl V) do otevřeného .xls dokumentu

Automatizace reportu: architektura řešení



Automatizovaný report: pro kont. tabulky

Analysis of contingency tables summary report - Mozilla Firefox

Soubor Úpravy Zobrazení Historie Záložky Nástroje nápověda

file:///Data/nikola/Blaf/Doktor/MAT/Bayes/vy: Google

Analysis of contingency tables...

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

Hotovo

Automatizovaný report: frekvence

Frekvence - Mozilla Firefox

Soubor Úpravy Zobrazení Historie Záložky Nástroje Nápověda

file:///Data/nikola/Blaf/Doktor/Biostat/R/CrossTabs/vyst11.htm

Frekvence

Frekvence

A1Vek	počet	%
1	64	26,446
2	26	10,744
3	40	16,529
4	49	20,248
5	63	26,033
Celkem	242	100 %

A2Pracoviste.bla.fff	počet	%
1	66	27,273
2	142	58,678
3	34	14,05
Celkem	242	100 %

B2	počet	%
	209	86,364
denní stacionář	3	1,24
lůžně	9	3,719
por.sál	3	1,24
sál	17	7,025
záchranka	1	0,413
Celkem	242	100 %

A3Sidlo	počet	%
1	211	87,19
2	31	12,81
Celkem	242	100 %

B3	počet	%
----	-------	---

Hotovo

Automatizovaný report: kont. tab. 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

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	0,12	0,12	3,42	1,59	-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	0,13	0,24	0,24	-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	600	0	1	2	1	#	0	0	0,33	0,12	1	#	-1,05	-1,83	0,27	-1,42	1,67		
2	200	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	340	0	0	6	0	#	0,16	0	0	0	0	#	-0,28	-0,78	-1,36	-0,78	0,35	-0,45	
4	412	2	0	4	0	#	0,2	0,22	0	0,24	0	#	-0,61	0,15	-0,88	0,35	-0,5		
5	541	5	1	2	0	#	0,26	0,33	0,33	0,12	0	#	-0,17	0,29	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

Automatizovaný report: pro spojité proměnné

Korelace a testy normality - Mozilla Firefox

Soubor Úpravy Zobrazení Historie Záložky Nástroje nápověda

file:///Data/nikola/Blaf/Doktor/MAT/Bayes/vyst02.htm

Korelace a testy normality

Korelace a testy normality

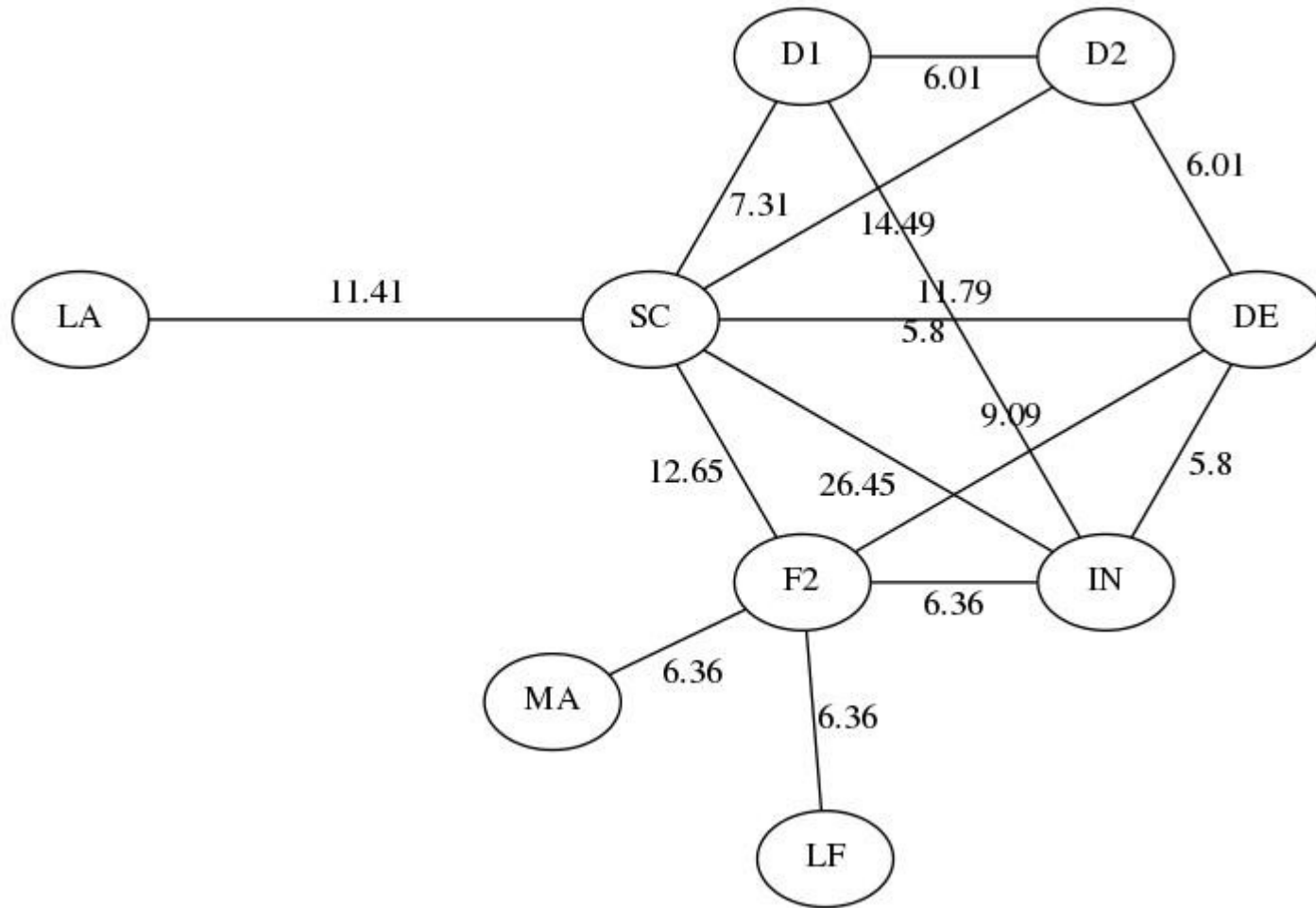
Normalita	N	p-hodnota	Kurt	p-hodnota	Skew	Min	Max	Průměr	Medián
LA	45	0,765	0,004	0	5	3,9	5		
MA	45	0,713	0,004	0	5	3,8	5		
LF	45	0,326	0,008	0	5	3,5	5		
D1	45	0,15	0,067	0	5	3,5	5		
D2	45	0,07	0,533	0	5	2,8	3		
F2	45	0,055	0,606	0	5	2,4	2		
IN	45	0,011	0,604	0	5	2,7	3		
DE	45	0,017	0,913	0	5	2,6	2,5		
SC	45	0,589	0,113	0	40	25,3	27		

Korelační koeficient # významnost # df

Pearson	LA	MA	LF	D1	D2	F2	IN	DE	SC	#	LA	MA	LF	D1	D2	F2	IN	DE	SC	#	LA
LA	1	0,367	0,229	0,286	0,387	0,126	0,214	0,361	0,538	#	0	0,013	0,131	0,057	0,009	0,408	0,158	0,015	0	#	43
MA	0,367	1	0,235	0,29	0,226	0,36	0,251	0,251	0,543	#	0,013	0	0,12	0,054	0,136	0,015	0,096	0,096	0	#	43
LF	0,229	0,235	1	0,272	0,198	0,289	0,194	0,267	0,513	#	0,131	0,12	0	0,071	0,192	0,054	0,201	0,076	0	#	43
D1	0,286	0,29	0,272	1	0,485	0,408	0,54	0,545	0,731	#	0,057	0,054	0,071	0	0,001	0,005	0	0	0	#	43
D2	0,387	0,226	0,198	0,485	1	0,373	0,422	0,52	0,688	#	0,009	0,136	0,192	0,001	0	0,011	0,004	0	0	#	43
F2	0,126	0,36	0,289	0,408	0,373	1	0,482	0,543	0,696	#	0,408	0,015	0,054	0,005	0,011	0	0,001	0	0	#	43
IN	0,214	0,251	0,194	0,54	0,422	0,482	1	0,56	0,727	#	0,158	0,096	0,201	0	0,004	0,001	0	0	0	#	43
DE	0,361	0,251	0,267	0,545	0,52	0,543	0,56	1	0,789	#	0,015	0,096	0,076	0	0	0	0	0	0	#	43

Hotovo

Automatizovaný report: pro 2x2 tabulky - graf závislostí podle Fisherova testu, OR



...

E-mail: data@tulipany.cz
Web: <http://data.tulipany.cz>

