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“

zdroj: http://www.acm.org

odbočka: Za co se dá dostat ACM ocenění

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

zdroj: http://www.acm.org

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)

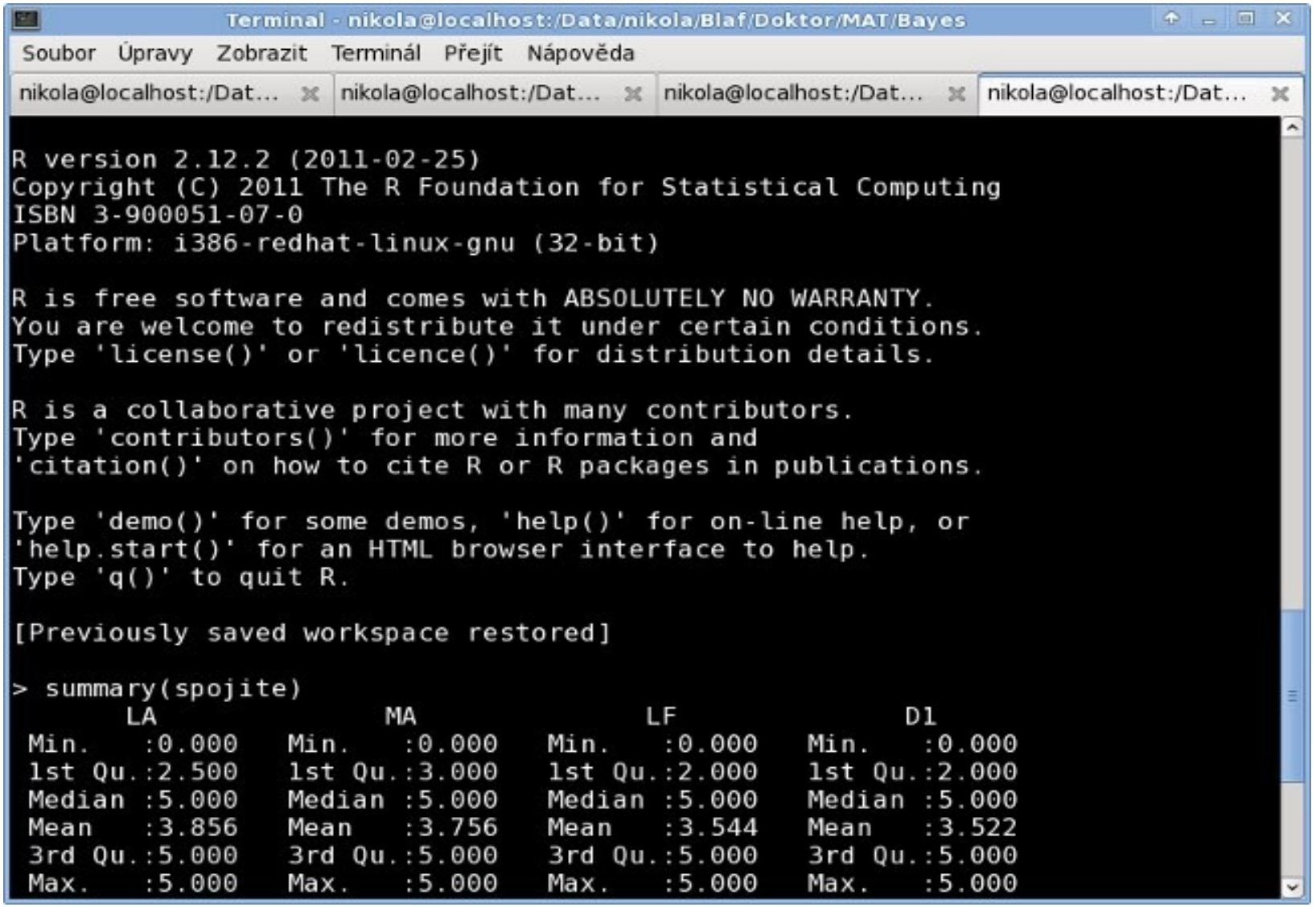
- Sada nástrojů Bioconductor – (od r. 2002) projekt vyvíjený s podporou Fred Hutchinson Cancer Research Center
 “an open source and open development software project for the analysis and comprehension of genomic data.“

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, ...

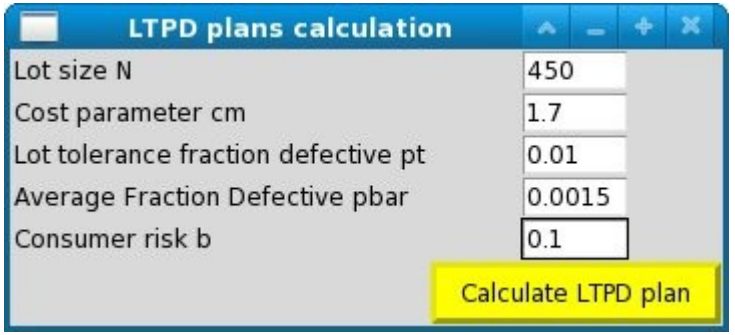
Jak může R vypadat?
... příkazová řádka



Jak může R vypadat / 2

... GUI vyvinuté na míru

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



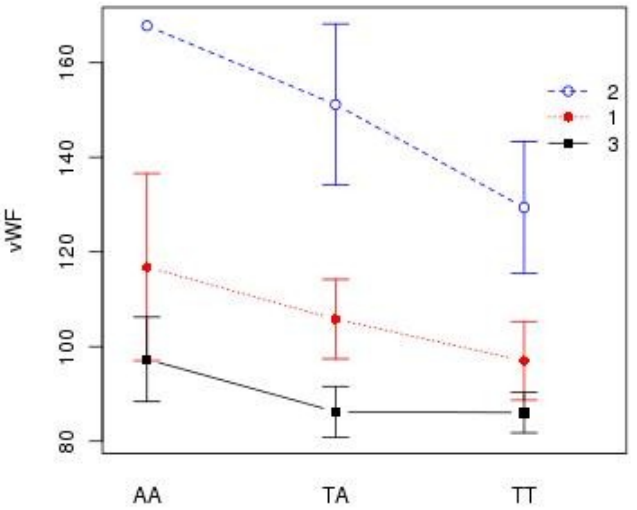
Grafické uživatelské rozhraní může uživateli odlehčit od nutnosti orientovat se v syntaxi ... ale ztratíme tím některé výhody

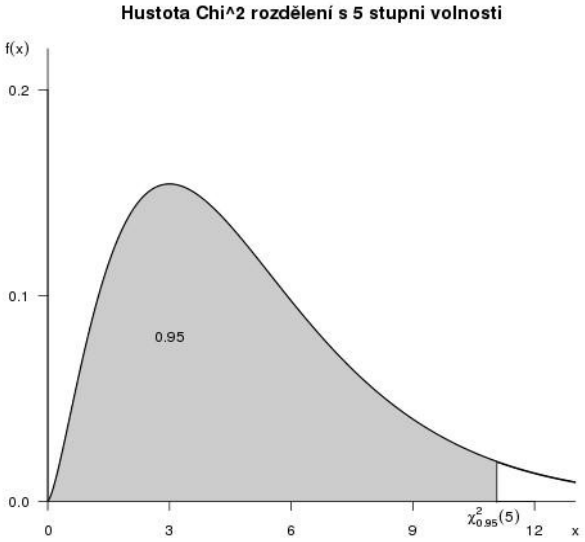
Grafika v R

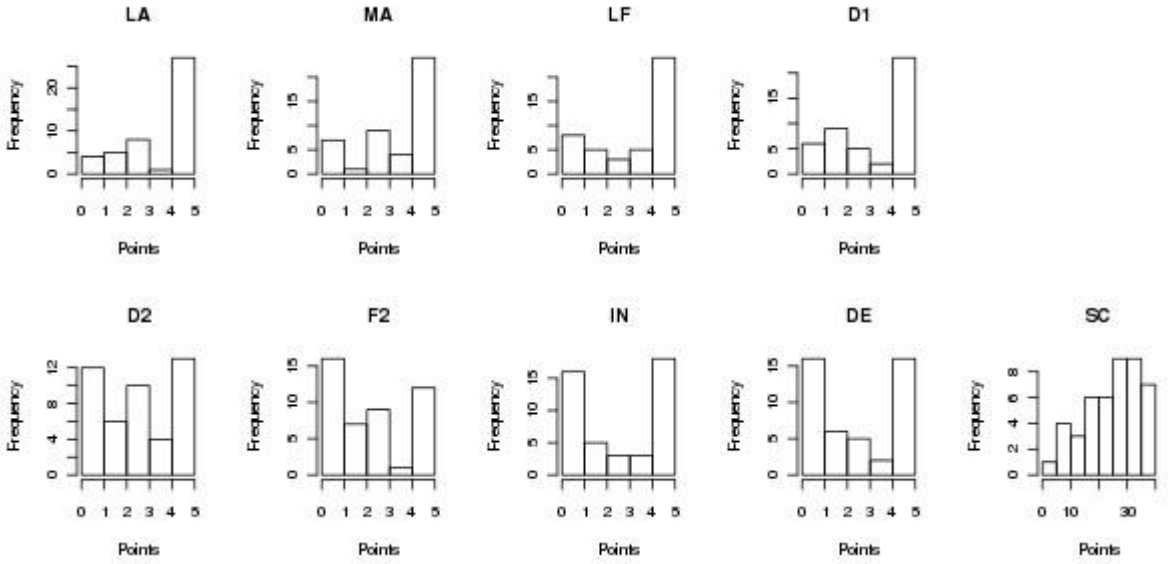
Tradiční systém, grid a lattice

Velice flexibilní

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







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**:

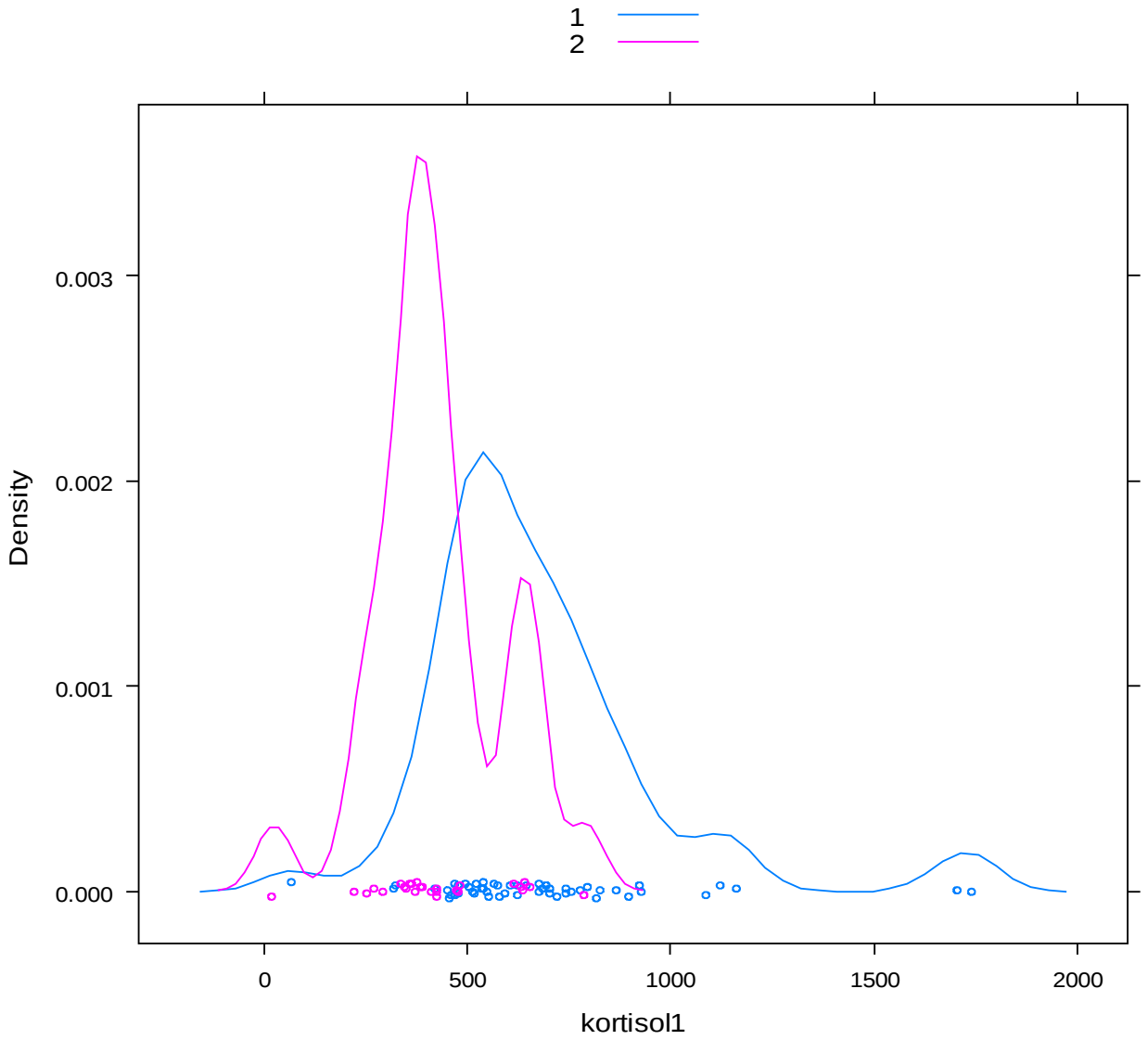
„Jsou v datech nějaké **zajímavé** vztahy?“

(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

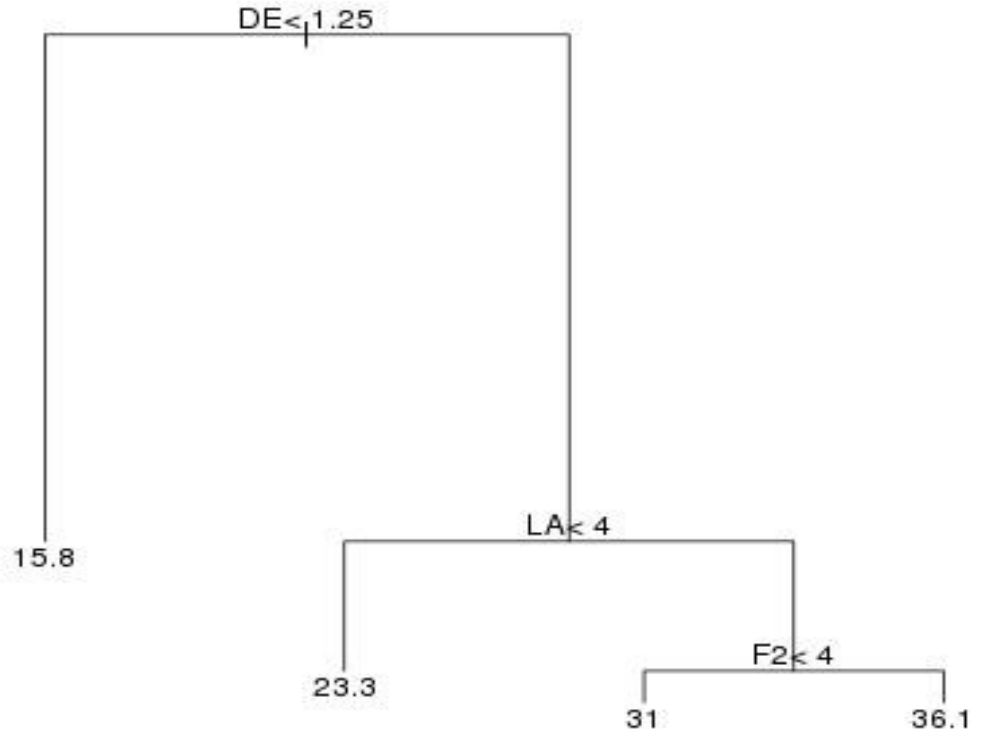
Možnost zachytit složitější interakce, vztahy platné jen pro určitou podskupinu

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

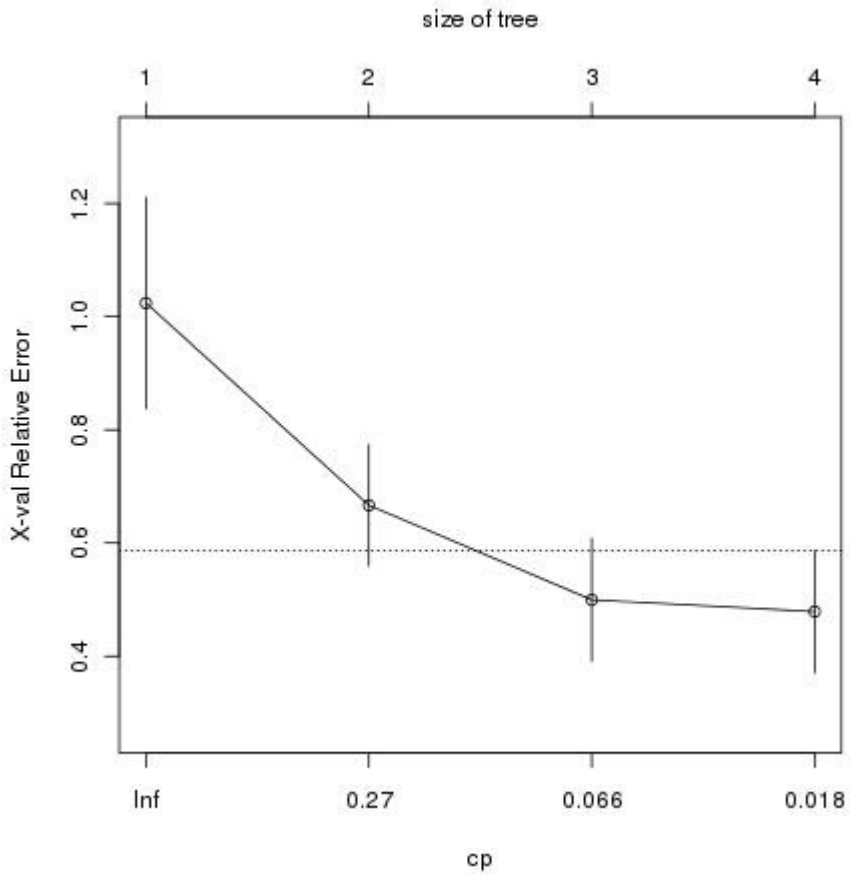
Výsledek modelování je (někdy) přehledný, snadná interpretace

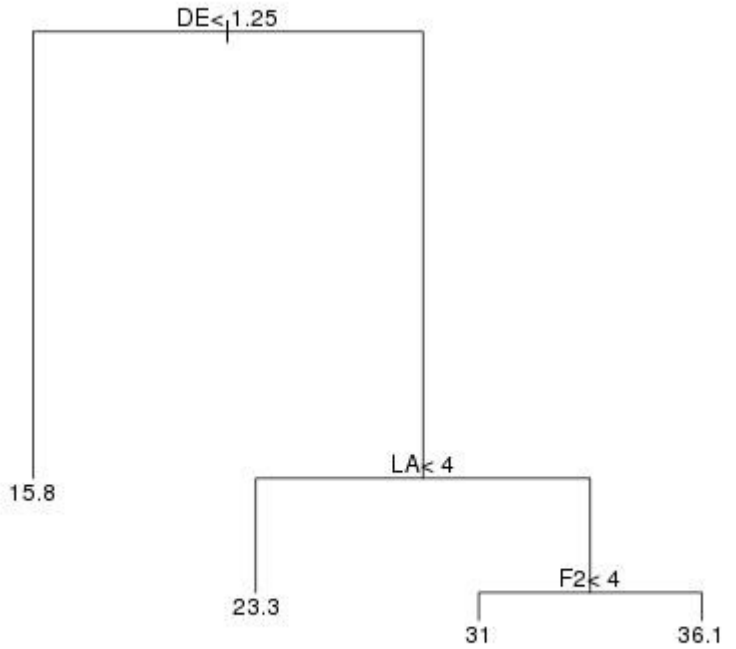
Použitelné pro identifikaci důležitých proměnných

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

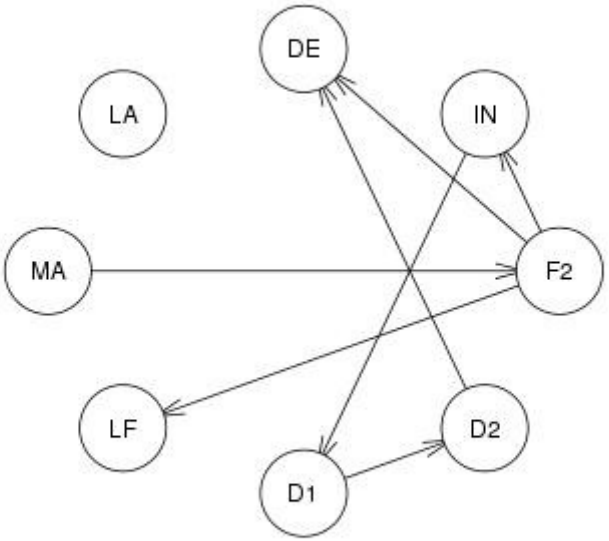
directed arcs: 7

learning algorithm: Hill-Climbing

score: Bayesian Information Criterion

penalization coefficient: 1.903331

tests used in the learning procedure: 392

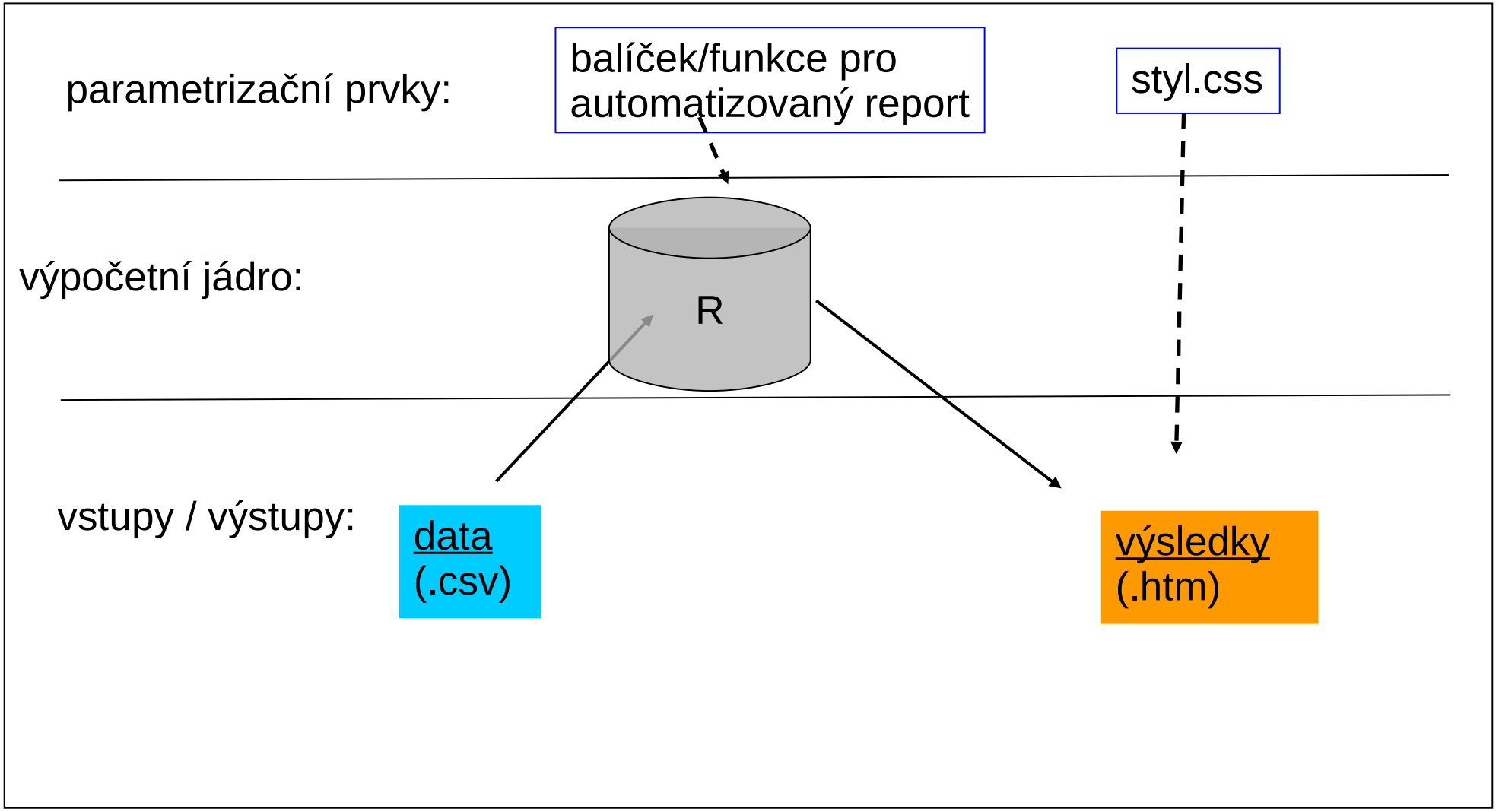


Automatizovaný reporting

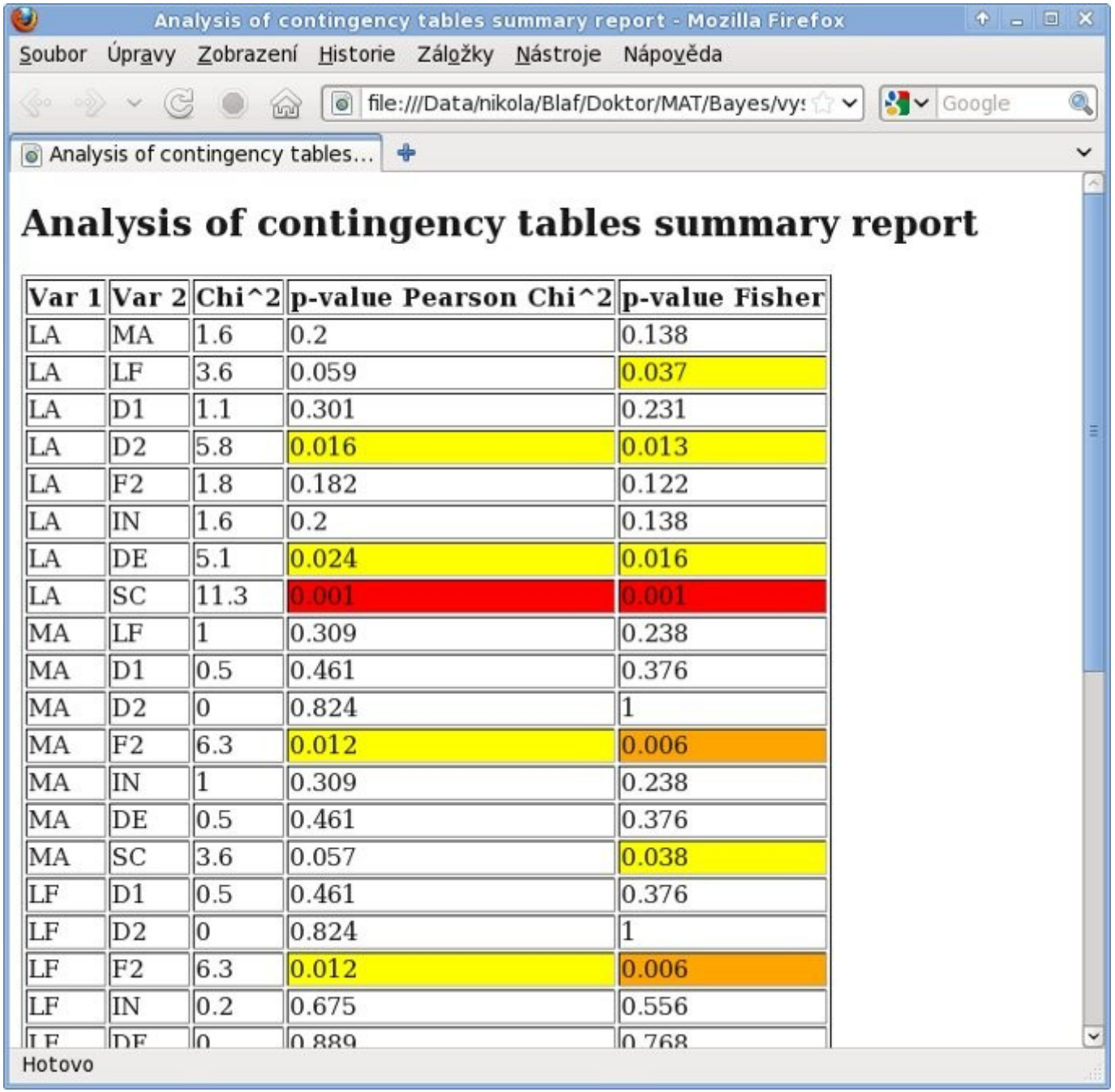
- arch. návrh založen na sledovaných principech
  - modularita – řešení je sadou jednotlivých komponent, které jsou poměrně nezávislé (-> přehledné a snadno přístupné úpravám), s jasně definovaným vstupem a výstupem
  - řešení postavené na otevřených standardech, nezávislé na platformě: využití formátu 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



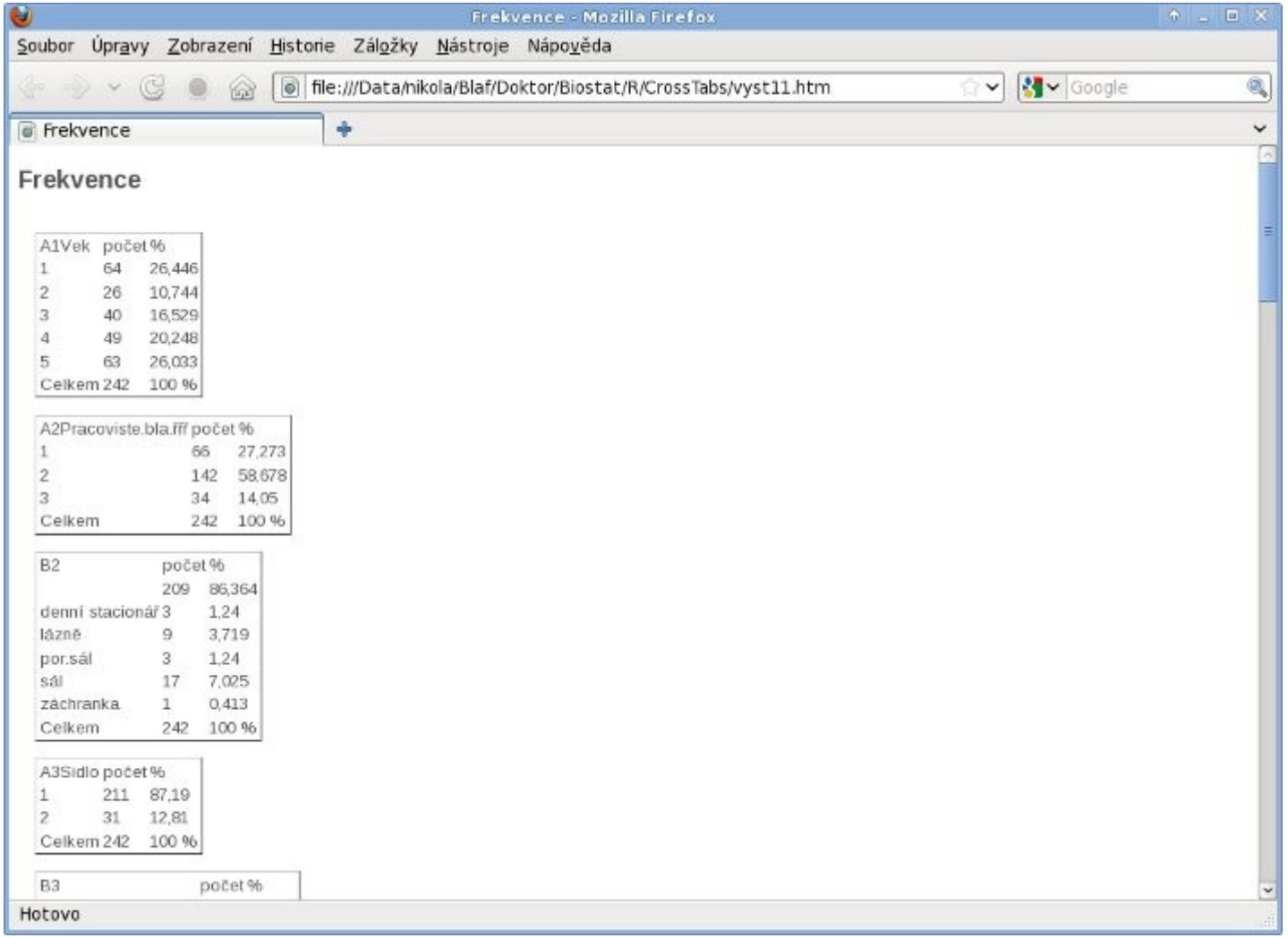
Automatizovaný report: pro kont. tabulky



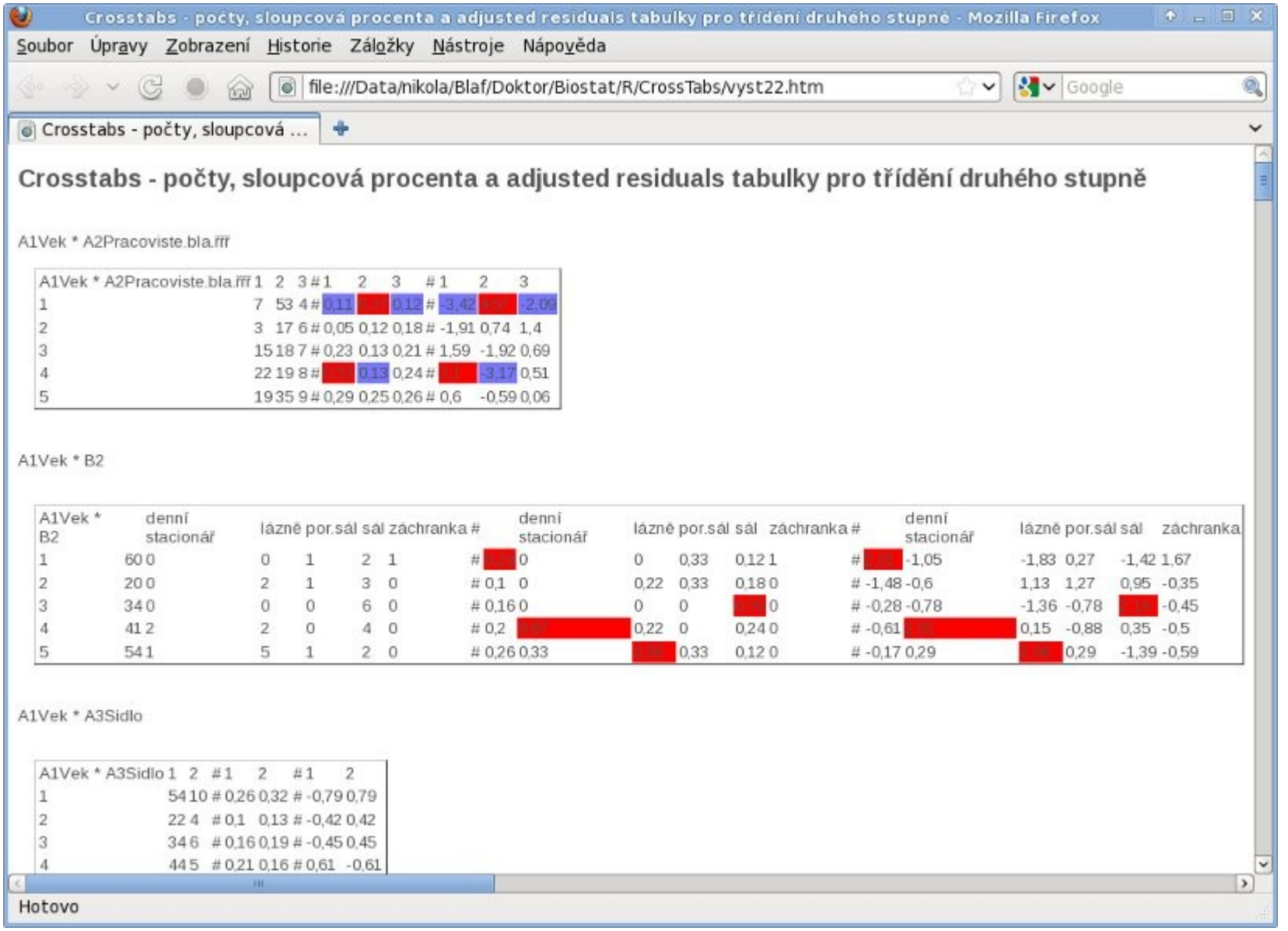
Analysis of contingency tables summary report

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

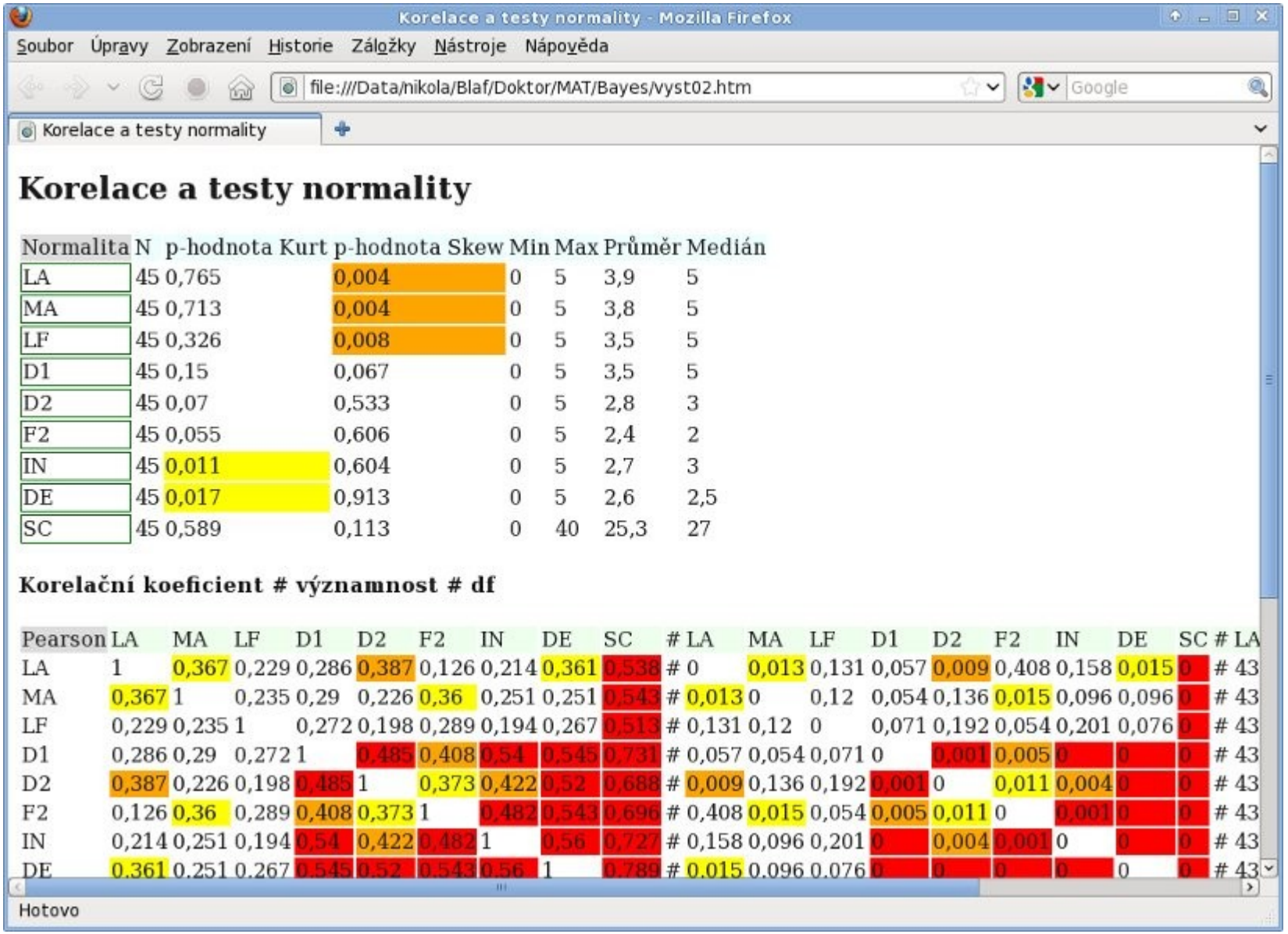
Automatizovaný report: frekvence



Automatizovaný report: kont. tab. 2



Automatizovaný report: pro spojité proměnné



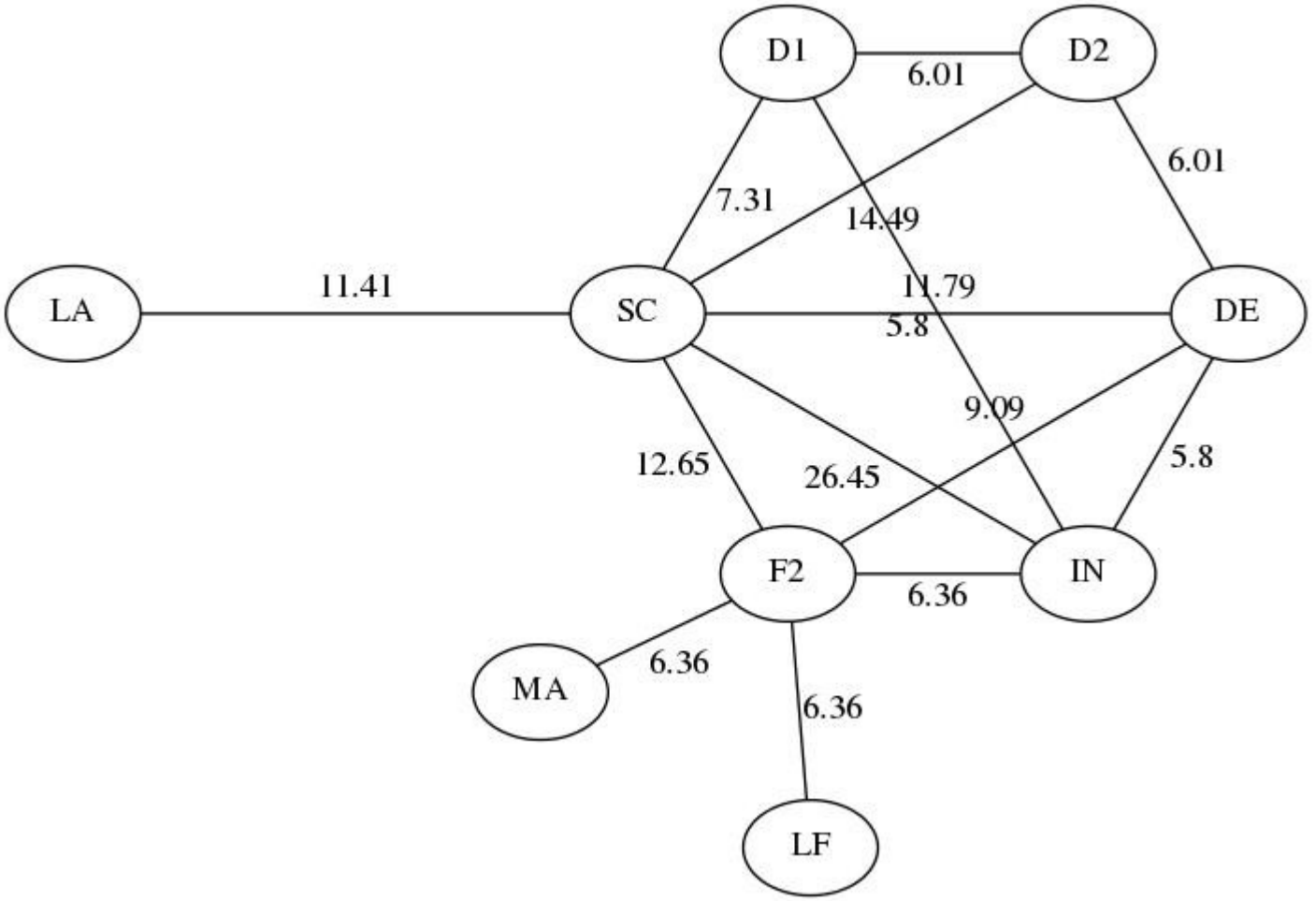
Korelace a testy normality

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

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



...

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