

## SYSTÉM R JAKO NÁSTROJ STATISTICKÉ ANALÝZY V BIOMEDICÍNSKÉM VÝZKUMU

Jan Hendl

### Souhrn

Příspěvek pojednává některé aspekty využívání programů pro statistickou analýzu. Ukazuje se, že vedle tradičních programových systémů jako SPSS nebo STATISTICA se stále více prosazuje systém R se svým programovacím jazykem. Tento systém má dvě hlavní výhody. Je volně šiřitelný stejně jako programové balíčky, které jsou v něm vytvářeny. Existuje již několik tisíc programových balíčků v jazyce R pro nejrůznější úkoly statistické analýzy. Pojednáváme základy systému R a některé rozšiřující programové balíčky.

### Klíčová slova:

*software, statistické programy, SPSS, jazyk R*

### Úvod

V lékařském výzkumu je statistická analýza dat běžným postupem. Bez vhodného programu bychom nemohli analyzovat efektivně naše data. V dnešním světě softwaru zaměřeného na statistickou analýzu dat se setkáváme se značnou varietou různých přístupů. Kromě velkých profesionálních statistických systémů s dlouhou tradicí jako SPSS, SYSTAT, STATISTICA nebo SAS, existují menší profesionální systémy (StatGraphics, NCSS) nebo volně dostupné systémy (OPENSTAT, FastStat, atd.). Statistické výpočty lze také provádět a programovat v matematicky orientovaných systémech jako například MATLAB. V minulém století i dnes má statistik samozřejmě možnost si vytvářet vlastní programy v jazycích jako FORTRAN nebo PASCAL či C. K tomu má k dispozici knihovny předprogramovaných statistických procedur v těchto jazycích. V současné době je velmi populární vytvářet nové aplikace v jazyce R, který byl navržen speciálně pro tvorbu statistických programů. Jeho obliba stále roste. Náš příspěvek se věnuje základnímu popisu prostředí a jazyka R. Dále popisujeme tři důležité systémy, které usnadňují práci v R. Jedná se o rozhraní k systému R: R – **Commnader**, **RStudio** a implementaci R do prostředí tabulkového procesoru MS Excel (**RExcel**).

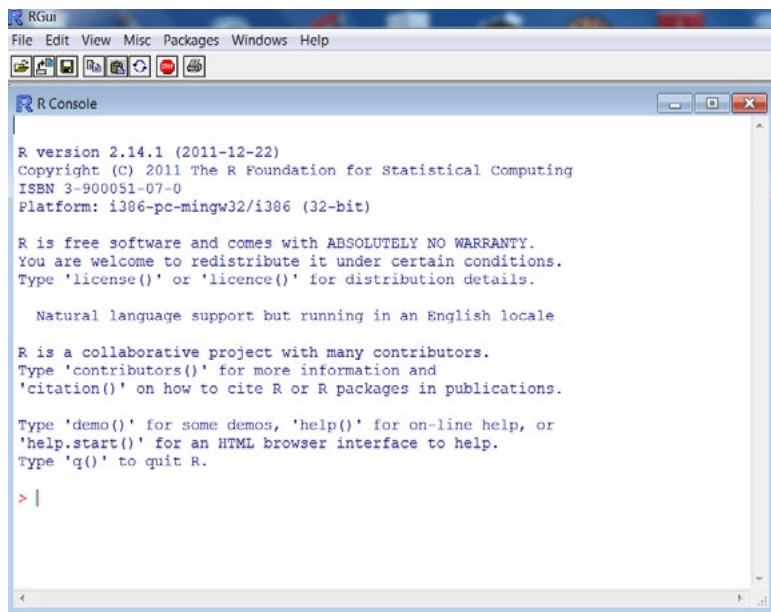
### Statistický systém a jazyk R

R je jazyk a prostředí pro statistické výpočty a jejich grafickou interpretaci. Informační stránky o projektu včetně volně šiřitelného softwaru nalezneme na adrese: <http://www.r-project.org/>. Z R se stal standard mezi statistiky. Systém R se využívá stále častěji ve výuce statistiky a je mnohem uživatelsky příjemnější než jiné programovací jazyky využívané pro návrh statistických systémů (např. C++ nebo Fortran). Má značnou podobnost s komerčním S-Plus jazykem a prostředím, vyvinutým v Bellových laboratořích. Obsahuje mohutné statistické makropříkazy a možnosti další příkazy a procedury naprogramovat v jazyku R.

Kromě jeho nulové ceny existuje mnoho důvodů, proč systém R získává na oblibě:

- R je velmi dobře navržený software s komerční kvalitou. Jeho vývojem se zabývali významní odborníci na statistické výpočty (D. Bates, J. Chambers, B. Ripley a další).
- R jazyk je dobře navržen s ohledem jak na praktikující statistiky tak na počítačové experty. Například jeho autor John Chambers získal cenu společnosti ACM (Ass. for Computing Machinery).
- Má výbornou podporu statistické grafiky.
- Protože zohledňuje požadavky na dobrý programovací jazyk, lze systém R relativně jednoduše rozšiřovat a doplňovat.
- R systém je možné instalovat na mnoha počítačových platformách, včetně Windows, Unix/Linux, Macintosh.

Na rozdíl od klasických statistických systémů, které využívají klikací grafické rozhraní pro své ovládání, R systém v původní podobě je příkazově orientovaný. Uživatelé řídí vykonávání operací pomocí příkazů v jazyce R. Interpreter příkazů v jazyce R odpovídá interaktivně na tyto příkazy. Je pravda, že R poskytuje určité možnosti běžného grafického rozhraní, ale ty jsou velmi skromné. Například můžeme pomocí něho instalovat programový balíček (viz dále). R rozhraní má také možnosti pro editování příkazů ve formě skriptů. Na obrázku 1 vidíme podobu standardního rozhraní systému R po spuštění.

The image shows a screenshot of the R GUI. At the top is a menu bar with 'File', 'Edit', 'View', 'Misc', 'Packages', 'Windows', and 'Help'. Below the menu bar is a toolbar with several icons. The main window is titled 'R Console' and contains the following text:

```
R version 2.14.1 (2011-12-22)
Copyright (C) 2011 The R Foundation for Statistical Computing
ISBN 3-900051-07-0
Platform: i386-pc-mingw32/i386 (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.

Natural language support but running in an English locale

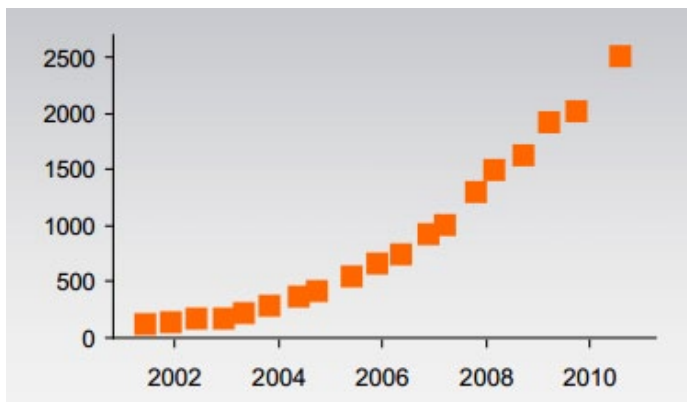
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

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.

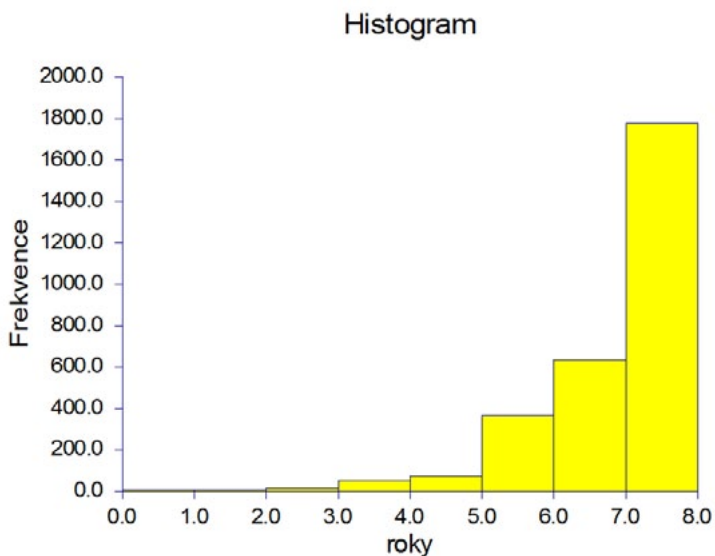
> |
```

Obrázek 4 — Rozhraní systému R se základní nabídkou nahoře a příkazovým řádkem začínajícím symbolem >.

Systém R lze používat v jeho původní podobě, ale stále více se uplatňuje využívání volně dostupných speciálních programových balíčků (packages) v jazyce R, které se vyvolávají v prostředí R. Tyto balíčky usnadňují mnoho standardních statistických operací a práci s daty. V dalším výkladu si všimneme balíčků R – Commander, RStudio. Kromě nich existuje několik tisíc balíčků, které implementují další operace a hlavně různé statistické procedury pro nejrůznější úkoly a situace vznikajících při analýze dat. Rostoucí dynamiku



Obrázek 2 — Vývoj počtů R balíčků od roku 2001 do roku 2011



Obrázek 3 — Četnosti změn v databázi R balíčků mezi roky 2004 a 2011

tvorby R balíčků ukazuje obrázek 2. Do konce roku 2011 bylo vytvořeno celkem 2950 balíčků. V období 2003–2007 to bylo pouze 82 balíčků. Počty změn v databázi balíčků v závislosti na čase v prvních 8 letech od vzniku systému R (do roku 2011) zobrazuje obrázek 3. Poznamenejme, že v lednu 2012 bylo zaregistrováno v databázi 420 změn balíčků (většina změn znamená dodání zcela nového balíčku). Uvedená data poukazují na oblibu R systému a vytváření R balíčků praktikujícími statistiky a dalšími výzkumníky, kteří chtějí své nápady zpřístupnit svým kolegům po celém světě. R balíčky jsou samozřejmě vytvářeny i odborníky z oborů mimo statistiku.

Tabulka 1 ukazuje výpis z přírůstků do databáze R balíčků poslední den měsíce prosinec v roce 2011.

30.12.2011	CDVine	Statistical inference of C- and D-vine copulas
30.12.2011	CellularAutomaton	One-Dimensional Cellular Automata
30.12.2011	cin	Causal Inference for Neuroscience
30.12.2011	coxme	Mixed Effects Cox Models
30.12.2011	fame	Interface for FAME time series database
30.12.2011	fast	Implementation of the Fourier Amplitude Sensitivity Test (FAST)
30.12.2011	genridge	Generalized ridge trace plots for ridge regression
30.12.2011	languageR	Data sets and functions with "Analyzing Linguistic Data: A practical introduction to statistics"
30.12.2011	MigClim	Simulating migration under climate change
30.12.2011	NormalGamma	Normal-gamma convolution model
30.12.2011	oro.nifti	Rigorous - NIFTI+ANALYZE+AFNI Input / Output
30.12.2011	R2OpenBUGS	Running OpenBUGS from R
30.12.2011	random.polychor.pa	A Parallel Analysis With Polychoric Correlation Matrices
30.12.2011	RcppExamples	Examples using Rcpp to interface R and C++
30.12.2011	ror	Robust Ordinal Regression MCDA library
30.12.2011	RSAGA	SAGA Geoprocessing and Terrain Analysis in R
30.12.2011	RTisean	R interface to Tisean algorithms
30.12.2011	testthat	Testthat code. Tools to make testing fun :)
30.12.2011	tis	Time Indexes and Time Indexed Series
30.12.2011	wasim	Visualisation and analysis of output files of the hydrological model WASIM

Tabulka 1 — Charakteristiky (datum dodání, název s hypertextovým odkazem, charakteristika) R balíčků pro den 30.12.2011 v databázi projektu R.

Standardní příkazově orientované rozhraní k statistickému softwaru (implementované v systému R) poskytuje určité výhody, zvláště:

- Je jednodušší v něm opravovat, modifikovat nebo opakovat analýzu. Mnoho statistických systémů (např. SPSS) mají grafické rozhraní umožňující generovat příkazy, je však zřejmé, uživatelé, kteří se spoléhají na grafické rozhraní tuto možnost nevyužívají, preferují klikání.
- Je jednodušší a přirozenější uchovávat záznam své práce, včetně vstupních parametrů a výstupů. R poskytuje mnoho triků pro sofistikované programování, kdy můžeme přirozeným způsobem příkazy, výstupy a osvětlující text (package).
- Grafická rozhraní statistických systémů jsou složitá a přesto neúplně pro pružné určení úkolu statistické analýzy.
- Jestliže přecházíme od grafické klikací nabídky k příkazově orientovanému systému, přináší to však určitý diskomfort. Přímé učení jazyku R bez předchozí zkušenosti s klikacími systémy je možná příhodnější. R povzbuzuje hlubší porozumění statistickým procedurám, než třeba SPSS. Na druhou stranu také pro systém R jsou k dispozici balíčky – jak jsme zmínili – simulující možnost klikání (např. R — Commander, který bude popsán dále).

V tomto příspěvku nechceme popsat jazyk R podrobně. Pouze poskytneme základní vstupní informace. R je funkcionální programovací jazyk, programy implementované v jazyce R mají formu funkcí a výrazy v jazyce R vyvolávají funkce. Například operátory pro sčítání + nebo násobení \* implicitně vyvolají funkci.

Možnosti jazyka R jsou rozsáhlé, lze zadávat stovky různých příkazů. Po spuštění se otevře hlavní okno (obr. 1) a v něm menší okno příkazové konzole (R Console). Na horní liště hlavního okna je několik základních příkazů spouštěných kliknutím přes nabídku nebo pomocí tlačítka.

Po úvodním textu (viz obr. 1) se objeví prompt > (znak větší než). Ten signalizuje místo, kam se píše výrazy jazyka R. Interpreter pak provede výrazy zapsané na řádce a zakončené klávesou ENTER. Uvedeme příklady:

```
> 1 + 2*3
[1] 7
> 1:3
[1] 1 2 3
> 1:3 + c(2, 5, -1)
[1] 3 7 2
> x <- 1:3
> 5*x
[1] 5 10 15
```

Většina ze zobrazených výstupů a vstupů jsou pochopitelné, některé však potřebují vysvětlení:

- Jednoduchý výstup začíná symbolem [1]. Jestliže výstup má více řádků, pak na začátku řádky se objeví její číslo podobně jako u prvního řádku.
- c() je funkce kombinování, která vytvoří vektor z jejích parametrů,

proměnných. Obecně argumenty jsou mezi závorkami a odděleny čárkami, jestliže argumenty mají jméno (name), pak jim přiřadíme hodnotu (value) takto: name = value.

- <- je přiřazovací operátor, v příkladu přiřazujeme hodnoty 1, 2, 3 proměnné x. Pro přiřazení lze však také použít znak =. Proměnné jsou vytvářeny dynamicky, jejich jména sestávají z kombinace velkých a malých písmen, číslic a dalších znaků, nemohou začít číslicí nebo podtržítkem \_. Rozlišujeme velká a malá písmena.
- Všechny objekty v příkladu jsou vektory. Kromě vektorů, používáme skaláry, datové rámce (data frame), tabulky (tables) a seznamy (list). R také rozlišuje objekty generované pomocí objektového programování.

Nyní vytvoříme vektor s jménem mujvektor s hodnotami 8, 6, 9, 10 a 5, musíme na příkazový řádek napsat:

```
> mujvektor<- c(8, 6, 9, 10, 5)
```

Abychom viděli co mujvektor obsahuje, napíšeme jeho jméno:

```
> mujvektor
[1] 8 6 9 10 5
```

Můžeme extrahovat jakýkoliv prvek tohoto vektoru

```
> mujvektor[4]
[1] 10
```

Na rozdíl od vektoru seznam (list) obsahuje prvky různých typů, numerické nebo znakové hodnoty, objekt list může jako prvek obsahovat také vektor. Funkce list() použijeme k vytvoření seznamu. Například vytvoříme seznam mujseznam takto:

```
> mujseznam <- list(jmeno="Franta", zena="Marie",
mujvektor)
```

Obsah seznamu zobrazíme uvedením jeho jména:

```
> mujseznam
$jmeno
[1] „Franta“
$zena
[1] „Marie“
[[3]]
[1] 8 6 9 10 5
```

Prvky seznamu jsou očíslovány a lze na ně odkázat číslicí v dvojitéch hranatých závorkách. Například:

```
> mujseznam[2]]
[1] „Marie“
> mujseznam[[3]]
[1] 8 6 9 10 5
```

Prvky seznamu mohou být také pojmenovány a pak vyvolány jménem seznamu

a jménem položky oddělené znakem \$. Například `mujseznam$jmeno` je totéž jako `mujseznam[[1]]` a `mujseznam$zena` je totéž jako `mujseznam[[2]]`:

```
> mujseznam$zena
[1] „Marie“
```

Můžeme nalézt jména pojmenovaných prvků použitím funkce `attributes()`: Například:

```
> attributes(mujseznam)
$names
[1] „jmeno“ „zena“
```

Jestliže použijeme funkci `attributes()` k nalezení jména prvků seznamu, jména jsou vždy uvedena za symbolem `“$names”`.

Jiným typem jazyka R je proměnná tabulka (`table`). Například jestliže vytvoříme vektorovou proměnnou `mojejmena` obsahující jména dětí ve třídě, použijeme funkci `table()`:

```
> mojejmena <- c(„Marie“, „Jan“, „Anna“, „Jirka“,
„Ota“, „Marie“, „Jarda“, „Jan“, „Franta“)
> table(mojejmena)
mojejmena
Anna Franta Jan Jarda Jirka Marie Ota
1 1 2 1 1 2 1
```

Můžeme vytvořit proměnnou `mojetabulka` a v ní uchovat obsah tabulky:

```
> mojetabulka <- table(mojejmena)
```

Opět k přístupu k jednotlivým hodnotám tabulky použijeme dvojité hranaté závorky:

```
> mojetabulka[[3]]
[1] 2
```

Kromě toho můžeme použít k získání třetího prvku tabulky jeho jméno:

```
> mojetabulka[,„Jan"]
[1] 2
```

Datový rámeč (`frame`) je důležitá datová struktura v R. Lze si jí představit jako dvourozměrné pole (resp. matici), přičemž jednotlivé sloupce nemusí být tvořeny vektory stejného typu. Datový rámeč je ve své podstatě vlastně tabulka, ve které se lze pomocí indexů odkazovat na jednotlivé buňky. Datové rámeče jsou velmi užitečné při zpracování dat, která jsme zapsali například do sešitu v tabulkovém procesoru (MS Excel, OpenOffice.org Calc apod.). Nejjednodušší je exportovat tabulku z těchto systémů jako text, kde položky jsou odděleny tabelátorem (viz tab. 2). Vzhledem k tomu, že R akceptuje jako oddělovač desetinných míst desetinnou tečku, je nutné převést desetinné čárky na desetinné tečky nejlépe v nějakém obyčejném textovém editoru.

Máme-li nachystaný textový soubor pojmenovaný např. `vysledky.txt`.

Jeho import do podoby datového rámce provedeme pomocí příkazu `read.table` následujícím způsobem:

```
> mojetabulka <- read.table(„vysledky.txt“, header =
T, sep = „\t“)
```

Volba `header = T` říká, že každý sloupec v textovém souboru má jméno, které má být také použito pro název sloupce v datovém rámci. Volba `sep = „\t“` říká, že importovaná data ve zdrojovém souboru jsou oddělena pomocí znaku tabulátor.

Funkce v R vyžadují obvykle argumenty, což jsou proměnné nebo objekty, na kterých funkce provede nějakou operaci:

```
> log10(100)
[1] 2
```

V systému R získáme nápovědu funkcí `help()`. Například chceme nápovědu pro funkci `log10()`:

```
> help(„log10“)
```

Jestliže tuto funkci použijeme, objeví se rámeček s informací o zvolené funkci.

Jestliže si nejsme jisti jménem, ale známe jeho úsek, použijeme funkce `help.search()` nebo `RSiteSearch()`. Funkce `help.search()` hledá již instalovanou funkci (i v instalovaných balíčcích). Funkce `RSiteSearch()` hledá všechny R funkce (i v neinstalovaných balíčcích).

Například chceme nalézt funkci počítající směrodatnou odchylku, budeme ji hledat jménem „deviation“:

```
> help.search(„deviation“)
Odpověď systému je následující:
Help files with alias or concept or title matching
'deviation' using fuzzy matching:
genefilter::rowSds
Row variance and standard deviation of
a numeric array
nlme::pooledSD Extract Pooled Standard Deviation
stats::mad Median Absolute Deviation
stats::sd Standard Deviation
vsn::meanSdPlot Plot row standard deviations versus
row
```

Mezi nalezenými funkcemi je funkce `sd()` balíčku „stats“ (R balíček obsažený v standardní instalaci R), který využijeme pro výpočet směrodatné odchylky.

V uvedeném příkladu funkce `help.search()` nalezne relevantní funkci (zde `sd()`). Jestliže nenalezneme, co chceme, použijeme funkci `RSiteSearch()` k hledání funkcí popsanych na WWW stránkách projektu:

```
> RSiteSearch(„deviation“)
```

Výsledkem hledání funkcí `RSiteSearch()` budou vyhledané funkce a odkazy na příslušné diskusní příspěvky.



Lze také provádět výpočty s objekty jako skaláry nebo vektory. Například aritmetický průměr pro vektor mujvektor (z čísel 8, 6, 9, 10, 5) vypočítáme pomocí funkce `mean()`:

```
> mean(mujvektor)
[1] 7.6
```

Použili jsme vnitřní R funkce (další příklady: `length()`, `print()`, `plot()`, atd.). Vytváříme také vlastní R funkce pro výpočty, které použijeme pro různé opakující se objekty

Například vytvoříme funkci, která vypočte součet čísla 20 + druhou mocninu vstupního čísla:

```
> mojefunkce <- function(x) { return(20 + (x*x)) }
```

Příkaz `return()` vrátí vypočtenou hodnotu.

Jakmile jste funkci vytvořili, je přístupná v dalších výpočtech. Například použijeme navrženou funkci postupně pro dvě čísla 10 a 25:

```
> mojefunkce(10)
[1] 120
> mojefunkce(25)
[1] 645
```

Užitečná je také vnitřní funkce `summary()`, který vrací souhrn výsledků v závislosti na typu objektu, který ji předáme jako argument. Pokud jako argument zadáme numerický vektor, funkce zobrazí základní statistické veličiny, které popisují soubor. V následujícím příkladu si vytvoříme numerický vektor `a`, který bude obsahovat náhodně vygenerované hodnoty, které pochází ze souboru s normálním rozdělením. Následně vyzkoušíme příkaz `summary()`.

```
> a <- rnorm(2000)
> summary(a)
Min. 1st Qu. Median Mean 3rd Qu. Max.
-3.477000 -0.676300 0.004711 0.004043 0.681600
3.658000
```

Zakončení programu R provedeme příkazem:

```
> q()
```

## R – Commander

R – **Commander** (balíček `Rcmdr`) je volně šiřitelný program. Byl vyvinut prof. Johnem Foxem jako grafické uživatelské prostředí pro zjednodušení práce v R. Má nabídku s volbami pro spuštění různých statistických procedur. V tomto smyslu je přitažlivou alternativou k programům jako je NCSS, SPSS nebo STATISTIKA. Program R – Commander je především vhodný pro začátečníky a účastníky kurzů základů statistiky. Lze v něm kontrolovat správné vytváření příkazů v jazyce R. Je možné ho obohatit o další části (tzv. plug-ins), které rozšiřují možnosti tohoto programku. Uvádíme některé z nich:

- `RcmdrPlugin.Export` — Grafický export do jazyka LATEX nebo HTML

- *RcmdrPlugin.IPSUR* — Úvod do statistiky použitím jazyka R
- *RcmdrPlugin.SurvivalT* — R – Commander plugin pro analýzu dat o přežívání
- *RcmdrPlugin.TeachingDemos* — R - Commander plugin s učebními demo procedurami
- *RcmdrPlugin.epack* — R – Commander plugin pro analýzu časových řad

Aby bylo možné R – Commander používat, musíme instalovat systém R a balíček *Rcmdr* ze stránek projektu R spolu se všem podřazenými balíčky, který R – Commander využívá. Upozorníme, že R – Commander vyžaduje SDI nastavení prostředí R (single-document interface).

### Výhody používání R – Commander

1. Užitečný pro začátečníky v jazyce R při vytváření základních grafů a vytváření statistických modelů.
2. Má okno se skriptem, výstupní okno a okno s historií, což umožňuje uživateli modifikaci skriptu. Například může změnit označení a parametry grafu.
3. Lze ho rozšířit o další R balíčky.
4. Rozhraní je jednoduché a dobře srozumitelné pro začínajícího uživatele.
5. Položky nabídek, které nejsou relevantní, nejsou plně označeny a upozorňují uživatele na potřebu úpravy dat.

### Relativní nedostatky programu R – Commander

1. Je určen pro zájemce o statistiku (původně studenty statistiky). Není kladen důraz na podrobnosti analýzy.
2. Má omezené možnosti vyhodnocování modelu z perspektivy firemních analytiků.

Není určen pro zájemce bez znalostí statistiky (např. uživatele s pouhým cílem dolovat informace z dat).

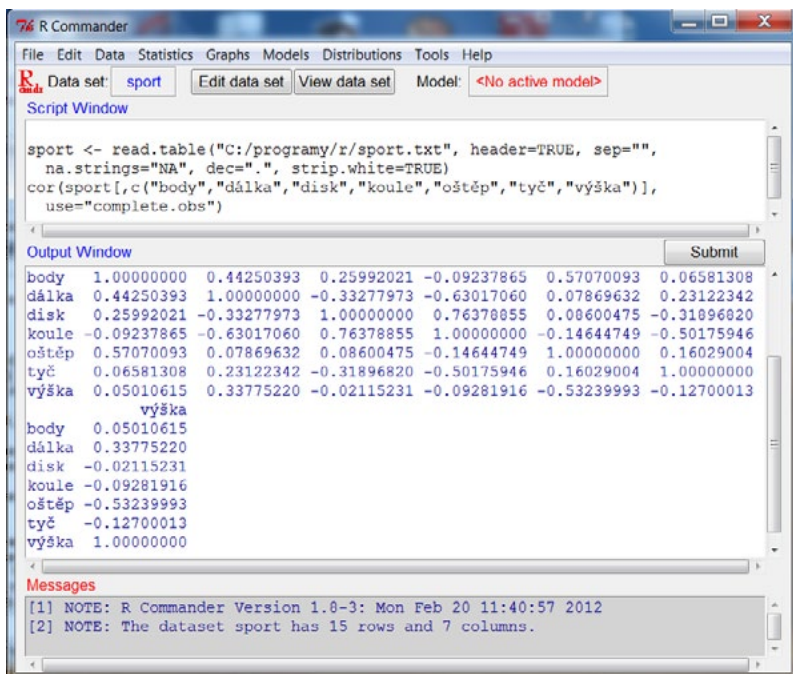
<b>Základní nabídky v R – Commander</b> <b>Rcmdr – jednotlivé skupiny jsou podobné nabídce v ostatních statistických programech</b>	
<b>File</b>	V této nabídce najdeme položky pro vyvolání a ukládání programů a skriptů v jazyce R a položky pro uložení výstupů výpočtů.
<b>Edit</b>	Obsahuje položky (Cut, Copy, Paste, atd.) pro práci s obsahem skriptů a obsahem výstupního okna s výsledky. Kliknutím pravým tlačítkem do skriptu nebo výstupního okna vyvolá nabídku EDIT. brings up an edit "context" menu
<b>Data</b>	Obsahuje položky pro čtení dat a manipulaci s daty.
<b>Statistics</b>	Obsahuje položky s běžnými statistickými procedurami.
<b>Graphs</b>	Obsahuje položky pro vytváření různých statistických grafů.
<b>Models</b>	Obsahuje speciální položky pro vyhodnocování statistických modelů.

Základní nabídky v R – Commander	
Rcmdr – jednotlivé skupiny jsou podobné nabídce v ostatních statistických programech	
<b>Tools</b>	Osbahuje položky pro vyvolání R balíčků nemajících vztah k Rcmdr (například pro nahrání dat s jiných statistických programů) a nastavení určitých voleb.
<b>Help</b>	Slouží k získání nápovědy o možnostech programu R – Commander.

V nabídce R – Commander jsou také tlačítka pro rychlé vyvolání akcí:

- **Data set:** ukáže jméno aktivního souboru dat a umožňuje přepnutí na jiný soubor, který je v paměti.
- **Edit data set:** umožňuje otevřít a editovat aktivní soubor dat.
- **View data set:** umožňuje prohlédnout aktivní soubor dat.
- **Model:** ukáže jméno aktivního statistického modelu a přepnutí mezi statistickými modely v paměti.

Demonstrujeme práci s R – Commander pomocí jednoduché analýzy dat s výkony atletů ve vybraných disciplínách desetiboje (obr. 4). V horním okně se znázorňují použité příkazy v jazyce R, které generuje R – Commander, v dolním okně jsou výsledky výpočtů..



Obrázek 4 — Okna systému R – Commander s hlavní nabídkou

Pomocí nabídky DATA→IMPORT jsme načetli data ze souboru sport.txt, který má podobu patrnou z tabulky 2.

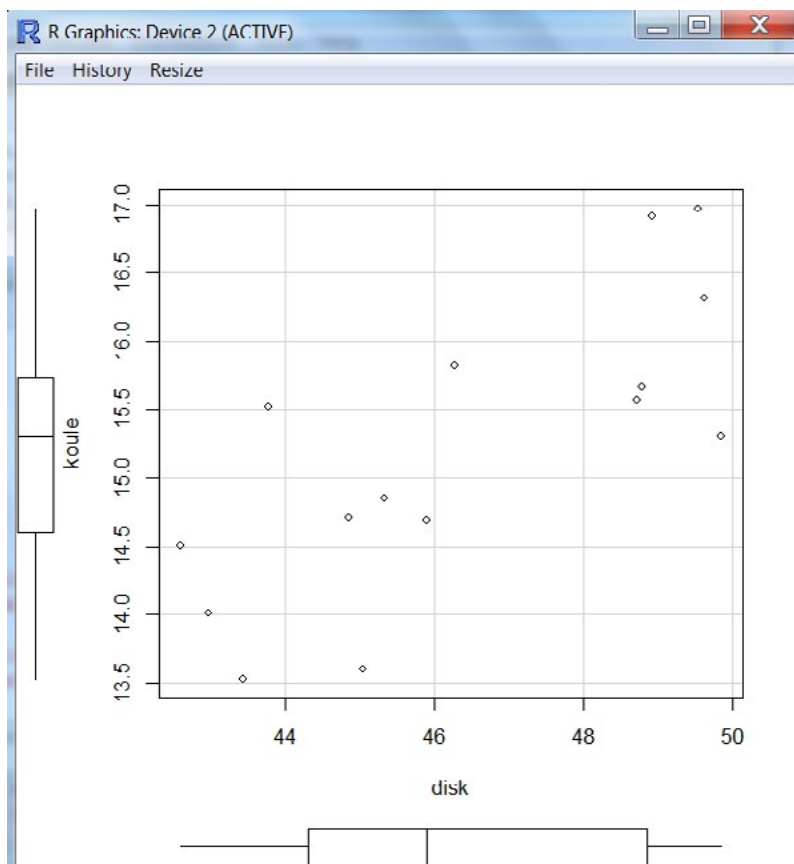
	"dálka"	"koule"	"výška"	"disk"	"tyč"	"oštěp"	"body"
"OBRIEN"	7.57	15.66	207	48.78	500	66.9	8824
"BUSEMANN"	8.07	13.6	204	45.04	480	66.86	8706
"DVORAK"	7.6	15.82	198	46.28	470	70.16	8664
"FRITZ"	7.77	15.31	204	49.84	510	65.7	8644
"HAMALAINEN"	7.48	16.32	198	49.62	500	57.66	8613
"NOOL"	7.88	14.01	201	42.98	540	65.48	8543
"ZMELIK"	7.64	13.53	195	43.44	540	67.2	8422
"GANIYEV"	7.61	14.71	213	44.86	520	53.7	8318
"PENALVER"	7.27	16.91	207	48.92	470	57.08	8307
"HUFFINS"	7.49	15.57	204	48.72	470	60.62	8300
"PLAZIAT"	7.82	14.85	204	45.34	490	52.18	8282
"MAGNUSSON"	7.28	15.52	195	43.78	480	61.1	8274
"SMITH"	7.47	16.97	195	49.54	500	64.34	8271
"MUELLER"	7.25	14.69	195	45.9	510	66.1	8253
"CHMARA"	7.75	14.51	210	42.6	490	54.84	8249

Tabulka 2 — Textový soubor sport.txt s daty o výkonech atletů

Přechodem do nabídky STATISTICS→SUMMARIES→CORRELATIONS a volbou proměnných vypočítáme korelační matici. Použité příkazy jsou zobrazeny v horním okně programu.

V dalším kroku pomocí nabídky GRAPHICS jsem vytvořili dvojrozměrný bodový graf závislosti výkon v hodu koulí a diskem (obr. 5).

Dále se omezíme na vyjmenování nejdůležitějších statistických a grafických procedur implementovaných v R – Commander:



Obrázek 5 —Grafický výstup po závislost výkonu v hodu koulí a oštěpem u sledovaných atletů.

Statistické procedury	Statistické grafy
Sumarizace dat (četnosti, tabulka s průměry, směrodatnými odchylkami, atd., korelační matice, test normality dat)	Volby barev v grafu
Kontingenční tabulky dvojrozměrné a vícerozměrné	Indexový graf
Statistická inference o průměrech	Histogram, graf lodyhy a listu, krabičkový graf
Statistická inference o četnostech	Kvantilový komparativní graf

Statistické procedury	Statistické grafy
Statistická inference o rozptylech	Jednoduchý shlukový graf a matice shlukových grafů
Neparametrické testy	Čárový graf, graf průměrů, sloupkový graf, koláčový graf,
Mnohorozměrné metody (metoda hlavních komponent, explorační a konfirmační faktorová analýza, shluková analýza, realibilita škály)	3D modifikace grafů
Predikční modely (lineární regrese, zobecněná lineární regrese, multinomická logitová regrese, ordinální regresní model)	Uložení obrázků s grafy jako PDF soubor

Modul R – Commander obsahuje bohaté možnosti zobrazování hustot a frekvenčních funkcí spojitych a diskrétních pravděpodobnostních rozložení včetně možnosti z nich generovat příslušná náhodná čísla.

Data lze přímo importovat pomocí speciálních funkcí systému R – Commander ze souborů ve formě SPSS, EXCEL nebo jednoduchých textových souborů (viz obr. 6).

### RExcel

RExcel je přídatný modul (add-in) pro MS Excel. Podporuje práci se systémem R uvnitř Excelu.

Jeho hlavní možnosti jsou:

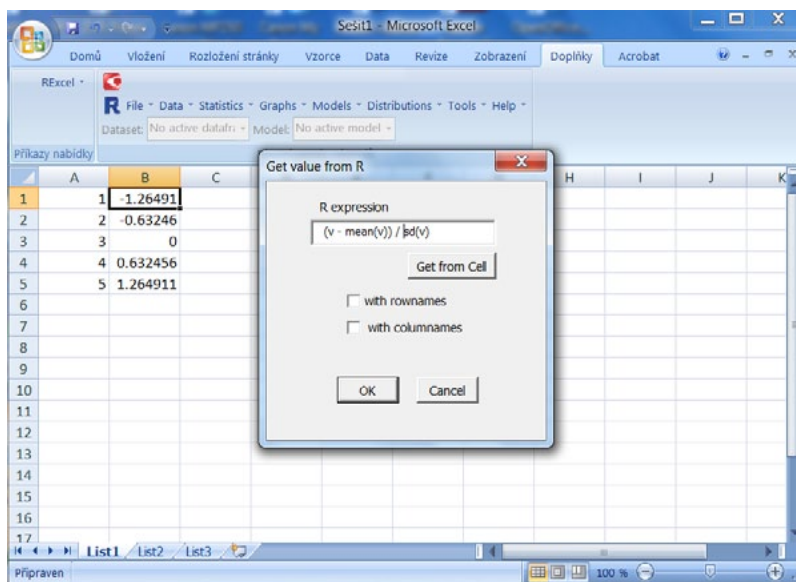
- Přenos dat mezi R a Excel oběma směry.
- Spuštění R skriptů přímo v Excelu
- Psaní maker, která volají R k provedení statistických výpočtů, aniž by by uživatel vnímal
- Volání R funkcí přímo ze vzorců v buňkách Excelu
- Využití Excelu jako rozhraní pro R, čímž umožňuje dostupnost funkcionality R pomocí nabídek a dialogových rámečků. Toto paradigma dovoluje využít také R Commander z nabídky Excelu.

Další informace o RExcel a instalační soubory nalezneme na adrese

<http://rcom.univie.ac.at/>.

Obrázek 6 přibližuje nainstalované prostředí RExcel spolu s modulem R – Commander a ukazuje, provedení standardizace pro vektor dat v sloupci A pomocí funkcí jazyka R.

Jestliže máme nainstalován systém R, pak lze instalovat systém RExcel jako balíček R. Následující množina kroků nahraje systém RExcel, konfiguruje RCOM server, instaluje RCOM server a nainstaluje RExcel instalační program. V systému musí být zpřístupněn balíček rcom, rsproxy a RExcelInstaller.



Obrázek 6 — Okno s RExcel v prostředí Excel s provedenou standardizací sloupce A pomocí funkce jazyka R.

Při instalaci celého systému postupujeme v těchto krocích:

```
> install.packages(„RExcelInstaller“, „rcom“,
„rsproxy“)
> # configure rcom
> library(rcom)
> comRegisterRegistry()
> library(RExcelInstaller)
> # provedeme následující příklad v R k provedení
příkazu v R pro start RDCOM
> installstatconnDCOM()
> # provedeme příkaz v R k nastartování instalace
RExcel
> installRExcel()
```

Řídíme se návodem instalace RExcel. Po instalaci RExcel, je možné vyvolat RExcel z nabídky programu EXCEL. Také se nainstalovala do Excelu nabídka balíčku R – Commander.

Jako jednoduchý test dokončení instalace provedeme následující úkoly:

Do sloupce A v Excelu napíšeme hodnoty 1 2 3 4 5. Dáme tyto hodnoty do bloku. V nabídce RExcel zvolíme položku „Put R Var“ > „Array“. Otevře se dialogové okno, vložíme jako jméno proměnné „v“ a stiskneme ENTER. Vytvoří se proměnná „v“ v systému R, která bude mít hodnoty ze sloupce A. Nyní přejdeme do buňky B1. V nabídce RExcel zvolíme položku „Get R Value“ > „Array“. Otevře se dialogové okno, vložíme výraz v jazyce R ( $v - \text{mean}(v) / \text{sd}(v)$ ) a stiskneme ENTER. Prove se standardizace hodnot vektoru v, výsledky se objeví v sloupci B (viz obr. 6).

### RStudio

RStudio je volně přístupný software, který rozšiřuje standardní rozhraní R. Stále v něm musíme psát příkazy, ale mnoho věcí se stává pohodlnější (Baier et al, 2011). Je určen především pro ladění programových skriptů. Abychom tento program nainstalovali, přejdeme na adresu <http://www.rstudio.org/> a stáhneme pomocí funkce Download RStudio ve verzi vhodné pro náš počítač.

Program RStudio je IDE systém (Integrated Development Environment). Jeho výhody jsou následující:

- RStudio usnadňuje interaktivní práci v systému R
- Data lze opravovat a formátovat pomocí několika komponent systému RStudio
- Editor v RStudio zlepšuje možnost opravovat kód příkazů a vytváření funkcí.
- Zjednodušuje hledání R balíčků, zrychluje vytváření vlastních balíčků.
- Vytváří pdf soubory s kombinací textu a výsledků výpočtů.
- RStudio poskytuje podporu „projektů“, které umožňují uživateli rychle přecházet mezi jednotlivými projekty. Každý projekt má vlastní pracovní adresáře, pracovní prostor a množinu zdrojových souborů.

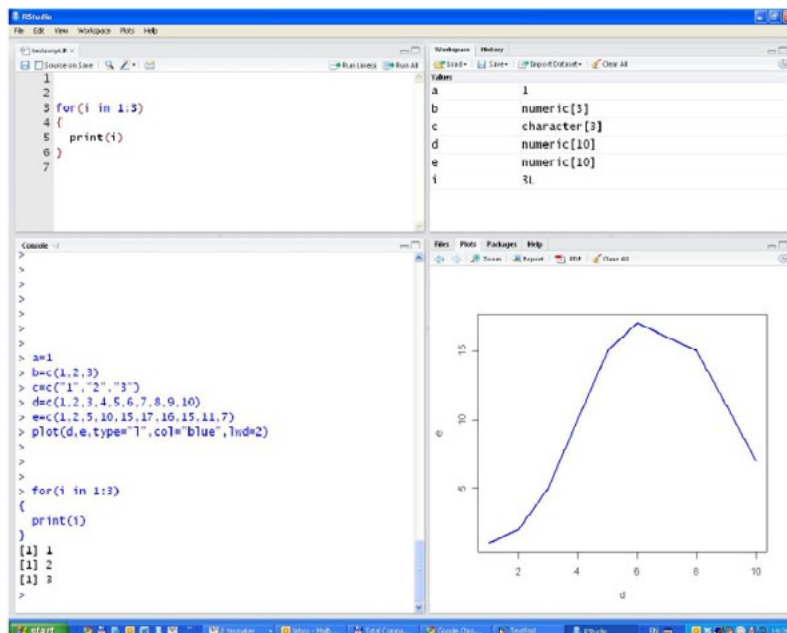
Rozhraní programu RStudio sestává z několika oken (viz. obr. 7)

Vlevo dole: okno konzole, příkazové okno. V něm vkládáme jednoduché příkazy jazyka R po výzvě >. Po stisknutí klávesy RETURN, systém R provede příslušný příkaz. Jedná se o nejdůležitější okno, protože v něm provádí R příslušné operace.

Vlevo nahoře: okno editoru, okno skriptování. Skript je množina příkazů a může být opravován a uložen na disk. Jestliže se toto okno neobjeví, lze vyvolat nabídkou FILE ->New ->R script

Vložením příkazů do tohoto okna nestačí pro vykonání příkazů. Je nutné je přenést do příkazového okna a v něm je provést. Jestliže to chceme provést, klikneme na nabídku Run nebo stiskneme CTRL+ENTER a tím přeneseme příkazy do příkazového okna.





Obrázek 7 — Program RStudio pro zjednodušení práce v systému R

Vpravo nahoře: workspace (pracovní prostor)/history (historie). V tomto okně uvidíme data a hodnoty systému R, které jsou v paměti. Je možné je opravit, jestliže na ně klikneme. Okno historie zobrazuje, co jsme vložili do systému a hodnoty.

Vpravo dole: files / plots / packages / help window. V tomto okně otevíráme soubory, prohlížíme grafy, instalujeme balíčky a zobrazujeme nápovědu.

V programu RStudio je možné měnit velikost popsaných oken. Veškerou práci v RStudio lze uchovávat v LATEX systému.

### Závěr:

Prostředí R se stalo standardem výpočetního softwaru pro statistiky a vzhledem k jednoduché dostupnosti a rozšiřitelnosti o speciální funkce je oblíben také u mnoha výzkumníků. Každý měsíc narůstá počet programových balíčků vytvořených v jazyce R o několik desítek. Každý z balíčků je určen pro jinou oblast statistické analýzy. V příspěvku jsme se podrobněji věnovali třem rozšířením a úpravám systému R, které umožňují využít systém při výuce statistiky a k přípravě statistických výpočtů v biomedicíně (srov. [Biomedicína]).

## Literatura:

- [1.] BAIER, T., NEUWIRTH, E., DE MEO, M. *Creating and Deploying an Application with (R)Excel and R. The R Journal* Vol. 3/2, December 2011.
- [2.] BÍNA, V., KOMÁREK, A., KOMÁRKOVÁ, L. *Jak na jazyk R: instalace a základní příkazy*. [online] 2006. 18 s. [cit. květen 2010]. Dostupné z WWW: <<http://www.karlin.mff.cuni.cz/~komarek/Rko/Rmanual2.pdf>>
- [3.] CHAMBERS, J. M. *Programming with Data*. Springer :New York, 1998.
- [4.] URL <http://cm.bell-labs.com/cm/ms/departments/sia/Sbook/>.
- [5.] EVERITT, B.S., HOTHORN, T. *A Handbook of Statistical Analysis Using R*. London : Chapman, Hall/CRC, 2006.
- [6.] [Biomedicína] Pdf soubor s popisem práce s R systémem při analýze biomedicinských dat: [https://github.com/avrilcoghlan/LittleBookofRBiomedicalStats/raw/master/\\_build/\\_latex/](https://github.com/avrilcoghlan/LittleBookofRBiomedicalStats/raw/master/_build/_latex/)

## Kontakt:

**Doc. RNDr. Jan Hendl, CSc.**  
Fakulta tělesné výchovy a sportu University  
Karlovy,  
katedra základů kinatropologie  
José Martího 31  
16252 Praha 6  
email: [hendl@ftvs.cuni.cz](mailto:hendl@ftvs.cuni.cz)