Vysoká škola ekonomická v Praze

Fakulta managementu v Jindřichově Hradci

Bakalářská práce

Vladimír Duchoň

Martina Slanařová

2007

Vysoká škola ekonomická v Praze

Fakulta managementu v Jindřichově Hradci

Katedra managementu informací

Možnosti R-commanderu při statistickém zpracování dat

Vypracovali:

Vladimír Duchoň

Martina Slanařová

Vedoucí diplomové práce:

Mgr. Komárková Lenka Ph.D.

Jindřichův Hradec, srpen 2007

Prohlášení

Prohlašujeme, že tuto bakalářskou práci na téma

» **Možnosti R-commanderu při statistickém zpracování dat** «

jsme vypracovali samostatně.

Použitou literaturu a podkladové materiály

uvádíme v přiloženém seznamu literatury.

.....

podpisy studentů

Anotace

Možnosti R-commanderu při statistickém zpracování dat

Cílem práce je prozkoumat možnosti nadstavby R-commander ve statistickém softwaru R. Výstupem práce by měl být popis, jak se pomocí tohoto nástroje dělají základní statistické analýzy a zdůraznění předností a nedostatků vůči klasickému R bez této nadstavby.

Poděkování

Na tomto místě bychom chtěli poděkovat vedoucí naší práce
paní Mgr. Lence Komárkové Ph.D. za poskytnuté rady a
připomínky, věnovaný čas a vstřícný přístup.

Obsah

1. Úvod

Možnosti R-commanderu při statistickém zpracování dat

Cílem naší práce je prozkoumání a zpřehlednění práce v R-commanderu, který je nadstavbou statistického programu R. Jelikož je tento program tzv. „freeware“, může být hojně využíván školami či jinými institucemi, které si nemohou dovolit vynakládat nemalé finanční částky potřebné při používání komerčních statistických softwarů. Program R, s již zmiňovanou nadstavbou R-commander, je vcelku jednoduchý na ovládání a na druhé straně umožňuje vypočítat a určit i složité statistické operace. Z těchto důvodů považujeme naši práci za přínosnou v této oblasti a doufáme, že tento jakýsi „návod“ ulehčí práci statistikům „laikům“ ještě více.

2. Instalace softwaru R

Dříve, než-li se pustíme do samotné práce s výše uvedeným statistickým programem, musíme jej nainstalovat. Tato operace není nikterak složitá, ale přesto si jí raději v této kapitole ukážeme.

Nejprve spustíme instalační soubor s názvem „R-2.5.0-win32.exe“, který nalezneme volně ke stažení na školním serveru *Klokan* a to v sekci „/PED/KMIH/verejny/kmm/Komarek/software/Rko-Windows“ nebo přímo z internetu na stránkách http://www.biometrics.mtu.edu/CRAN/bin/windows/base/. Po spuštění instalace si zvolíme jazyk, ve kterém bude instalace probíhat a pak už podle průvodce volíme možnost „Další“. Po výběru souhlasu s podmínkami licenční smlouvy a umístění instalovaného softwaru, se proces úspěšně dokončí.

2.1. Doinstalování knihoven

Již při spouštění instalačního souboru jsme si mohli povšimnout ještě jedné složky, která byla pojmenována „Baliky“. Tento adresář v sobě obsahuje balíčky s doplňujícími knihovnami, které bude program potřebovat při početních operacích, a proto je nezbytné tyto knihovny ještě doinstalovat. Jako první krok, který nyní uděláme, bude překopírování složky „Baliky“ někam na náš pevný disk v PC. Po otevření si můžeme povšimnout, že kromě souborů s příponou „ZIP“ je zde i jeden soubor nesoucí jméno „instal.R“. Tento soubor nám později bude sloužit jako jakýsi spouštěcí instalační mechanismus pro program R, a proto jej musíme nepatrně obměnit dle potřeb našeho PC. Editaci toho souboru můžeme velice jednoduše provést například stisknutím klávesy F4 (používáme-li Windows Commander, Total Commander apod.). Po otevření editačního okna musíme pozměnit první řádek tak, aby cesta, kterou tento řádek ukazuje, směrovala skutečně tam, kam jsme si uložili již zmiňovaný adresář „Baliky“ obsahující naše „zazipované“ knihovny. Při přepisování této cesty si musíme dát pozor především na to, abychom používali správná lomítka (/) a celou cestu tímto lomítkem také zakončili. V opačném případě by byla cesta neplatná.

Pokud máme editaci hotovou, uložíme změny a můžeme přistoupit k samotné instalaci knihoven. Ta probíhá již přímo z prostředí programu R, a proto ho nyní zpustíme. Do příkazového řádku zadáme: source("C:/xxx /yyy/zzz/www"), kde mezi lomítky bude postupně uvedena cesta vedoucí k námi editovanému souboru „instal.R“. Jako příklad zde můžeme uvést typickou cestu , a to: source("C:/Program Files/R//Baliky/install.R"). Tento zápis potvrdíme a tím se spustí i požadovaná instalace. Těmito kroky bychom měli mít software R připravený k dalšímu použití a to se všemi standardními knihovnami. Pokud bude zapotřebí nějaká knihovna navíc, tak jí obdobným způsobem doinstalujeme.

3. Úvod do práce s R-Comanderem

Vždy, když budeme chtít pracovat v prostředí R-commanderu, je nutné nejprve otevřít samotný program R, do jehož příkazového řádku zadáme příkaz „**library(Rcmdr)**“. Tento pokyn slouží pro otevření námi požadované nadstavby, která se objeví po potvrzení v novém okně.

Již na první pohled si můžeme povšimnout, že okno R-commanderu je rozděleno zhruba do 4 částí. Vrchní část tvoří Menu, pomocí něhož ovládáme celý R-commandr. Podrobněji se mu budeme věnovat v následují kapitole. Pod ovládacími prvky se nachází dvě okna. Horní z nich (Script window) slouží, jak už název vypovídá, k zapisování dat a příkazů. Jakoukoli operaci, kterou zadáme do „Script window, musíme potvrdit stiskem tlačítka „Submit“. Oproti tomu okno pojmenované Output window nám ukazuje výsledky operací, jejichž hodnoty nás zajímaly. Sem bude hlavně soustředěna naše pozornost. Místa, na kterých dané důležité hodnoty ve výsledcích najdeme, si ukážeme postupně na příkladech. Úplně dole, pod tímto oknem, si můžeme povšimnout tabulky s nápisem Messages. V tom dialogovém okně se nám budou zobrazovat chybová hlášení či jiné informace o povrzené operaci.V případě chybových hlášení je vhodné tato oznámení po odstranění nedostatku vymazat, abychom věděli, zda jsme skutečně předchozí omyl opravili či nikoli. Hlášení v této sekci jsou barevně rozlišena podle druhu informace. Chybová hlášení jsou zvýrazněna červenou barvou, varování jsou zelená a ostatní poznámky jsou psány modře. Při chybách a varováních je slyšet i zvukový signál, který nám dává najevo, že se vyskytla v procesu nějaká chyba.

4. Mapa menu

File

- Open script file…
- Save skript…
- Save script as…
- Save output…
- Save output as…
- Save R workspace…
- Save R workspace as…
- Exit
    - From Commander
    - From Commander and R

Edit

- Clear window
- Cut
- Copy
- Paste
- Delete
- Find…
- Select all

Data

- New data set…
- Import data
    - from text file…
    - from SPSS data set…
    - from Minitab data set…
    - from STATA data set…
- Data in packages
    - List data sets in packages
    - Read data set from attached package…

- Active data set
    - Select active data set…
    - Help on active data set (if available)
    - Variables in active data set
    - Set case names…
    - Subset active data set…
    - Remove cases with missing data…
    - Export active data set…
- Manage variables in active data set
    - Recode variable…
    - Compute new variable…
    - Standardize variables…
    - Convert numeric variable to factor…
    - Bin numeric variable…
    - Reorder factor levels…
    - Define contrasts for a factor…
    - Rename variables…
    - Delete variables from data set…

Statistics

- Summaries
    - Active data set
    - Numerical summaries…
    - Frequency distribution…
    - Table of statistics…
    - Correlation matrix…
- Contingency Tables
    - Two-way table…
    - Multi-way table…
    - Enter and analyze two-way table…
- Means
    - Single-sample t-test…
    - Independent-samples t-test…
    - Paired t-test…

        - One-way ANOVA…
        - Multi-way ANOVA…
- Proportions
        - Single-sample proportion test…
        - Two-sample proportions test…
- Variance
        - Two-variances F-test…
        - Bartlett's test…
        - Levene's test…
- Nonparametric tests
        - Two-sample Wilcoxon test…
        - Paired-samples Wilcoxon test…
        - Kruskal-Wallis test…
- Dimensional analysis
        - Scale reliability…
        - Principal-components analysis…
        - Factor analysis…
        - Cluster analysis
                - k-means cluster analysis…
                - Hierarchical cluster analysis…
                - Summarize hierarchical clustering…
                - Add hierarchical clustering to data set…
- Fit models
        - Linear regression…
        - Linear model…
        - Generalized linear model…
        - Multinomial logit model…
        - Proportional-odds logit model…

Graphs

- Index plot…
- Histogram…
- Stem-and-leaf display…
- Boxplot…

- Quantile-comparison plot…
- Scatterplot…
- Scatterplot matrix…
- Line graph…
- Plot of means…
- Bar graph…
- Pie chart…
- 3D graph
    - 3D scatterplot…
    - Identify observations with mouse
    - Save graph to file
- Save graph to file
    - as bitmap
    - as PDF/Postscript/EPS…
    - 3D RGL graph…

Models

- Select active model…
- Summarize model
- Add observation statistics to data…
- Confidence intervals…
- Hypothesis tests
    - ANOVA table
    - Compare two models…
    - Linear hypothesis…
- Numerical diagnostics
    - Variance-inflation factors
    - Breusch-Pagan test for heteroscedasticity…
    - Durbin-Watson test for autocorrelation…
    - RESET test for nonlinearity…
    - Bonferroni outlier test
- Graphs
    - Basic diagnostic plots
    - Residual quantile-comparison plot…

        - Component + residual plots
        - Added-variable plots
        - Influence plot
        - Effect plots

Distributions

- Normal distribution
        - Normal quantiles...
        - Normal probabilities...
        - Plot normal distribution...
- t distribution
        - t quantiles...
        - t probabilities...
        - Plot t distribution...
- Chi-squared distribution
        - Chi-squared quantiles...
        - Chi-squared probabilities...
        - Plot chi-squared distribution...
- F distribution
        - F quantiles...
        - F probabilities...
        - Plot F distribution...
- Binomial distribution
        - Binomial quantiles...
        - Binomial tail probabilities...
        - Binomial probabilities...
        - Plot binomial distribution...
- Poisson distribution
        - Poisson probabilities...
        - Plot Poisson distribution...

Tools

- Load package(s)
- Options

Help

- Commander help
- Introduction to the R Commander
- Help on active data set (if available)
- About Rcmdr

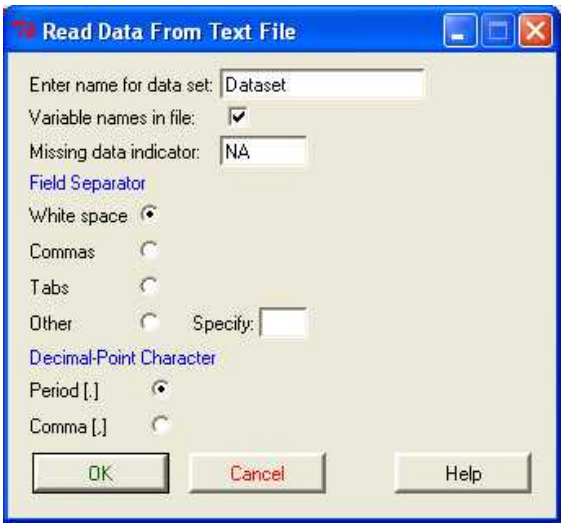
5. Data

5.1. Načtení dat

Statistická zkoumání nebo jiné modely vždy simulujeme na nějakém datovém souboru, a proto je jednou z nejdůležitějších operací, které si budeme muset vysvětlit, načtení dat. Této problematice se věnuje právě tato kapitola. Nadstavba R-commander nám umožňuje hned několik cest, jak datový soubor načíst.

5.1.1. Načtení dat ze souborů

Jako první je načtení dat z nějakého souboru. Za takovéto soubory můžeme požadovat soubory textové, SSPS, Minilab nebo soubory STATA. Pro každý typ je dána určitá koncovka, dle které lze soubory rozeznat. Nejčastěji bývají data zapsána v souboru textovém, jelikož aplikace „Poznámkový blok“ je součástí většiny OS. Takovéto soubory mají koncovku „txt“. Načtení dat je v tomto případě velice jednoduché a snadné. V základním menu R-commaderu klikneme na záložku *Data* a po té zvolíme *Import data*, dále položku *from text file*. Otevře se nám dialogové okno, (viz Obr.1) ve kterém si můžeme změnit jméno dat.



Obrázek 1. Načtení dat

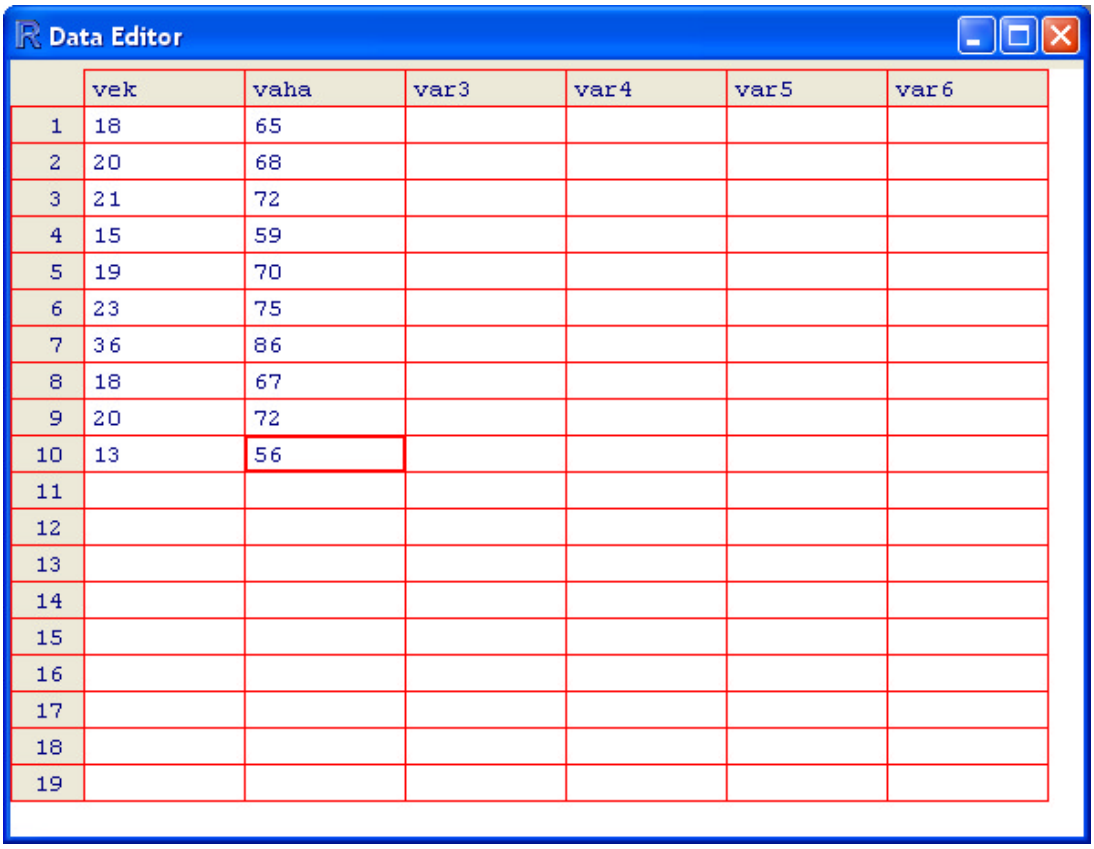
Standardně je přednastavený název na Dataset. Při přejmenování musíme dát pozor na psaní velkých a malých písmen, jelikož soubory dat s názvy například „data“ a „Data“ jsou dva rozdílné. Kromě názvu si můžeme ještě v témže dialogové okně povšimnout

řádku s nápisem *NA*. Toto políčko reprezentuje označení, jak je v otevíraném datovém souboru reprezentováno místo, kde nejsou data k dispozici. Obvykle se takto „nedostupná“ data nahrazují právě popisem NA(not available). Další „sekce“ v nastavení nám umožňuje zvolit zápisy jednotlivých hodnot v datovém souboru tak, aby byla data správně načtena. Můžeme zde volit mezi přednastavenými možnostmi a nebo zvolit nějakou specifickou, je-li to třeba. Nejčastěji používané oddělovače jednotlivých hodnot jsou zde přímo vypsány a můžeme tedy vybírat mezi mezerami, čárkami a tabelátory. Pokud by hodnoty byly odděleny nějak jinak, máme možnost tento znak vepsat do příslušné kolonky po zaškrtnutí možnosti *Others*. Jako poslední možnost nastavení zde vidíme volbu oddělení desetinných míst u jednotlivých hodnot. Na výběr je zde oddělení pomocí čárky, které je typické hlavně pro evropské státy a nebo oddělené tečkou, používané především v Americe. Po provedených úpravách stiskneme tlačítko OK. Následně se nám otevře další okno, které slouží již k samotnému nalezení a otevření datového souboru. Můžeme si povšimnout, že při této metodě načtení dat lze kromě souborů s příponou „txt“ otevřít i soubory typu „dat“. Požadovaný datový soubor načteme označením a stiskem tlačítka „Otevřít“. K tomu, abychom si mohli načtená data prohlédnout a ověřit, že se jedná doopravdy o správná data, slouží tlačítko „*View data set*“ v základní nabídce R-commanderu. Další možnost, jak zobrazit naše data, je pomocí vypsání názvu datového souboru do okna *Script Window* R-commanderu. Pokud jsme tedy naše data nepřejmenovali a název zůstal Dataset, napíšeme příkaz Dataset a potvrdíme kliknutím na Submit. Ihned si můžeme všimnout, že tabulka s daty se nám vypsala v okně Output Window. Ve spodní části R-commanderu Messages nám software kromě varovných hlášení zobrazuje také obecné informace, jak jsme si uvedli výše. V případě načítání dat se zde objevuje počet řádků a sloupců tabulky. Pokud bychom při zobrazení dat zjistili nějakou nesrovnalost nebo chybu, můžeme data opravit a to prostřednictvím tlačítka *Edit data* set, které je umístěno vedle funkce *View data set*. Pro možnost opravy dat se nám otevře nové okno, ve kterém máme možnost data editovat. Až dosáhneme požadované změny, okno editace zavřeme křížkem a data se nám automaticky uloží.

5.1.2. Přímé vkládání dat

Další možností jak „dostat“ data do R-commaderu je jejich přímé vložení. Možná by bylo přesnější napsat přímé vepsání, protože tato funkce nám umožňuje vytvoření

tabulky s daty v softwaru R, respektive v její nadstavbě. Ani tato metoda není nikterak složitá a slouží nám převážně v případě, kdy data nemáme v elektronické podobě a musíme si je teda do PC přepsat. V menu R-commanderu zvolíme záložku *Data* a poté vybereme možnost „*New data set...*“. V dalším kroku si volíme název datového souboru. Jako v předchozím případě je přednastaven název na *Dataset,* ale my si ho pro lepší přehlednost můžeme přejmenovat. Opět je zde třeba dát pozor na přesný název, aby nedošlo k záměně. Po potvrzení se nám otevře nové okno – Data editor. Do libovolného počtu řádku a sloupců můžeme nyní vepsat potřebné hodnoty. Pro změnu názvu sloupců nám postačí jedno „kliknutí“ na stávající název (defaultně nastavený na var1, var2,…,…). Až budeme mít tabulku sestavenou dle potřeby (Obr.2), okno editoru zavřeme křížkem. Hodnoty se stejně jako v předchozím případě uloží a my s nimi budeme moci dále pracovat. Pro vypsání nebo zobrazení tabulky použijeme stejný postup jako u obdobné operce při načítání dat ze souborů a to prostřednictvím tlačítka „*Wiev data set“* a nebo vypsáním názvu datového souboru a potvrzením pomocí příkazu „Submit“. R-commander nám i v tomto případě v kolonce Messages vypíše základní údaje o tabulce s daty.

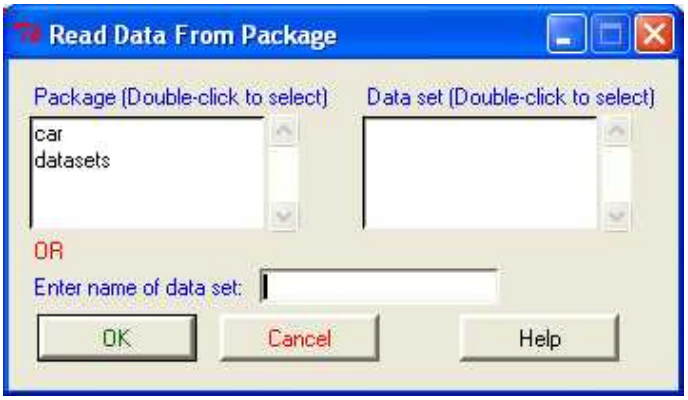


	vek	vaha	var3	var4	var5	var6
1	18	65				
2	20	68				
3	21	72				
4	15	59				
5	19	70				
6	23	75				
7	36	86				
8	18	67				
9	20	72				
10	13	56				
11						
12						
13						
14						
15						
16						
17						
18						
19						

Obrázek 2. Vkládání dat

5.1.3. Načtení dat z balíčku

Software R obsahuje různé balíčky, které v sobě skrývají uložená data. Jejich načtení můžeme provést pomocí příkazu Data v základním menu, když dále vybereme možnost „*Data in packages*“ a následně „*Read data set from an attaches package*…“. Objeví se nám tabulka (viz.Obr.3.). Zde si můžeme vybrat požadovaný balíček, který nám po „dvojkliku“ zobrazí další možnosti datového souboru. Pokud známe název požadovaného souboru, můžeme ho zadat přímo do kolonky „*Enter name of data set*“.



Obrázek 3. Vkládání dat z balíčků

5.2. Práce s datovým souborem

Data - Active data set - Select active data set

Zobrazí se nám datové soubory v paměti programu, z nichž si můžeme vybrat jeden soubor dat, se kterým můžeme dále pracovat.

Data - Active data set - Help on active data set (if available)

Otevře okno s nápovědou pro právě aktivní data. Nápověda obsahuje základní informace, například ze kterého balíku byla data načtena apod.

Data - Active data set - Variables in active data set

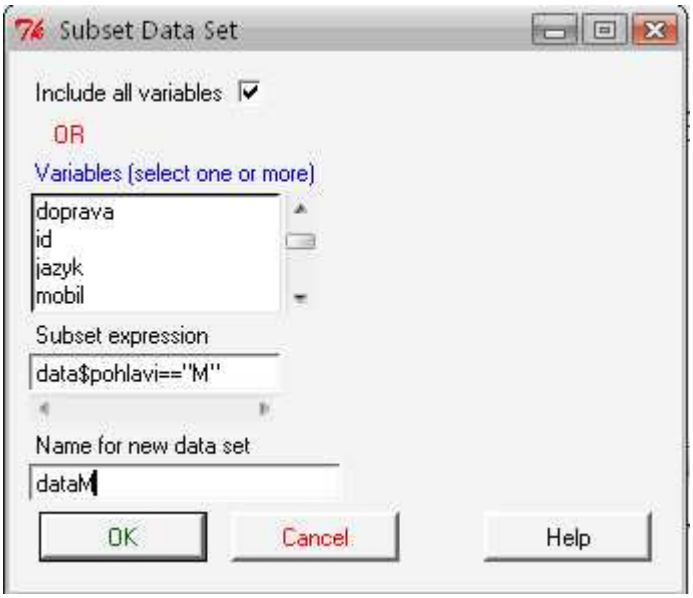
Vyjmenuje proměnné, které datový soubor obsahuje.

Data - Active data set - Set case names

Zde lze v nabídce označit ten sloupec, který slouží pouze k označení názvů jednotlivých řádků a ne jako samostatná proměnná. V našich datech k tomuto účelu slouží sloupec nazvaný „id“.

Data - Active data set – Subset active data set

Pomocí této funkce můžeme z dat doslova vybrat jejich určitou část, se kterou dále chceme pracovat. Pokud bychom například chtěli dále pracovat jen s daty pro pohlaví „M“, zadáme do kolonky **subset expression** *data$pohlavi = = „M“* a do kolonky **name for new data set** zadáme název nových, námi vytvořených dat, pro náš případ použijeme například *dataM*. (viz. Obr.4)



Obrázek 4. Vyplnění nabídky Subset

Tuto funkci můžeme také využít pro zúžení našich původních dat pouze na určité proměnné. K tomu stačí odznačit v horní části tabulky **include all variables** a po té vybrat v kolonce **varibles** proměnné, které potřebujeme.

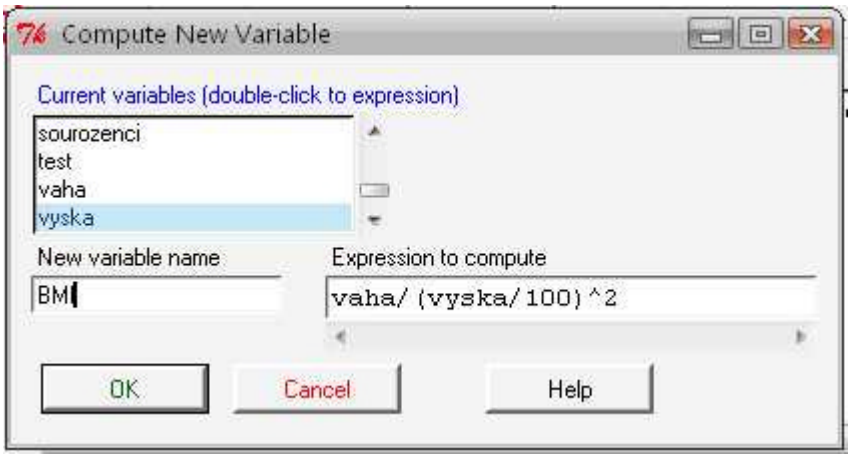
Data - Active data set - Remove cases with missing data
Odstraní případné chybějící hodnoty ve zvolené proměnné.

Data - Active data set - Export active data set
Uloží aktivní data do vybraného adresáře ve zvoleném formátu (.txt, .dat, .csv)

5.3. Práce s proměnnými

Data - Manage variables in active data set - Compute new variable
Tato funkce nám umožní doslova vytvořit novou proměnnou pomocí proměnných, které máme k dispozici.
Zkusme si například vypočítat novou proměnou „BMI“, a to pomocí proměnné váha a výška. (BMI se vypočítá jako váha v kg dělená výškou v metrech na druhou).

Do kolonky **Expression to compute** tedy zapíšeme potřebný vzorec a do kolonky **New variable name** jméno pro novou proměnnou (viz. Obr.5).



Obrázek 5. Vytvoření nové proměnné "BMI"

Data - Manage variables in active data set - Standardize variables

Provede se tzv. Z-transformace vybrané proměnné (tj. od jednotlivých hodnot se odečte průměr a poté se ještě vydělí směrodatnou odchylkou).

Data - Manage variables in active data set - Convert numeric variable to factor

Přemění numerickou proměnnou na kategoriální proměnnou.

Uveďme si jednoduchý příklad pro proměnnou sourozenci. Původně je tato proměnná chápána jako numerická proměnná. Po provedení příkazu **Summary** se nám proto spočítají jednotlivé charakteristiky (viz.Obr.6).

```
   sourozenci
 Min.   :0.0000
 1st Qu.:1.0000
 Median :1.0000
 Mean   :0.9516
 3rd Qu.:1.0000
 Max.   :2.0000
```

Obrázek 6. Sourozenci jako numerická proměnná

```
 sourozenci2
 0:11
 1:43
 2: 8
```

Obrázek 7. Sourozenci jako kategoriální proměnná

Po převedení této proměnné na faktor, nám funkce **Summary** vytvoří místo tabulky charakteristik tabulku četností, neboť už novou proměnnou chápe jako kategoriální, tj. 0,1,2 jsou pro něj jen symboly pro kategorie (viz. Obr.7.).

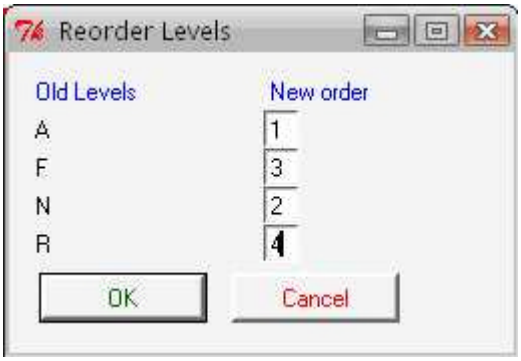
Data - Manage variables in active data set - Bin numeric variable

Rozdělí proměnnou na několik intervalů (počet intervalů se dá nastavit v kolonce **number of bins**). Můžeme si zde vybrat tři různé způsoby dělení. Na stejně velké intervaly (equal-width bins) nebo dělení na intervaly s přibližně stejným počtem (equal-count bins) nebo intervaly rozdělit přirozeně (natural breaks). Dále si zde můžeme zvolit označení intervalů. Pokud zvolíme možnost **Specify names,** můžeme si v nové tabulce vytvořit vlastní pojmenování jednotlivých intervalů. Zaškrtneme-li možnost **numbers**, pak se nám intervaly označí čísly. Pokud chceme, aby se intervaly pojmenovaly podle rozmezí, ve kterém se dané číslo nachází, zvolíme možnost **Ranges**.

Data - Manage variables in active data set - Reorder factor levels

Software, pokud není uvedeno jinak, řadí kategorie automaticky dle abecedy.

Například u proměnné jazyk je určeno pořadí A, F, N, R. Pokud potřebujeme docílit jiného pořadí, zvolíme právě tuto nabídku. Na obrázku číslo 8 například vidíme, jak lze změnit pořadí na A, N, F, R.



Obrázek 8. Změna pořadí na A, N, F, R

Tato nabídka dostane smysl hlavně u jiných proměnných jako je například „vzdělání“, původně program předurčuje abecední pořadí, tedy SS, VS, ZS, logicky je však správně ZS, SS, VS (v tomto případě je i dobré zatrhnout kolonku **Make ordered factor**).

Data - Manage variables in active data set - Rename variables
Umožní změnit názvy jednotlivých proměnných.

Data - Manage variables in active data set - Delete variables from data set
Vymaže námi zvolenou proměnnou.

6. Statistická analýza dat

6.1. Popisná statistika

Statistics - Summaries - Active data set

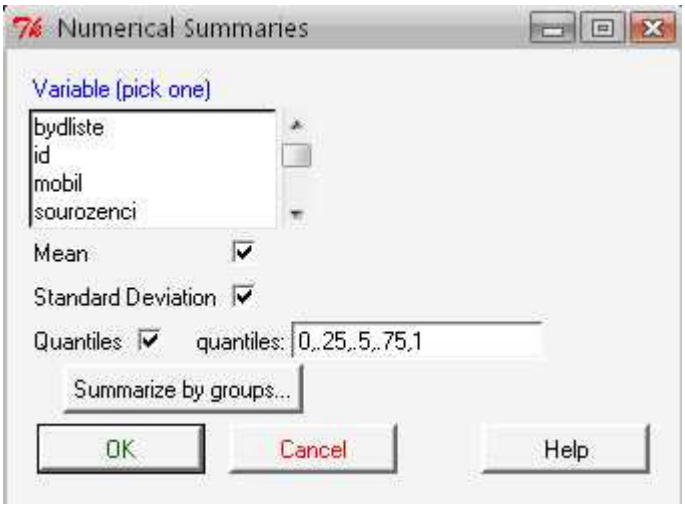
Jedna z nejzákladnějších funkcí, poskytne nám prvotní pohled na analyzovaná data. Pomocí této funkce, jak můžete vidět na obrázku číslo 9, jsme si pro každou kvantitativní proměnou spočítali průměr, medián, horní a dolní kvartil, minimum a maximum. Pro každou kategoriální proměnnou nám funkce spočítala četnosti jednotlivých znaků.

```
> summary(data)
       id            bydliste     doprava   sourozenci         mobil
 Min.   : 1.00   Min.   :  0.0   A: 2    Min.   :0.0000   Min.   :0.0000
 1st Qu.:16.25   1st Qu.:  0.0   B:19    1st Qu.:1.0000   1st Qu.:0.0000
 Median :31.50   Median :114.5   M: 9    Median :1.0000   Median :1.0000
 Mean   :31.50   Mean   :133.9   T: 7    Mean   :0.9516   Mean   :0.7258
 3rd Qu.:46.75   3rd Qu.:173.8   V:25    3rd Qu.:1.0000   3rd Qu.:1.0000
 Max.   :62.00   Max.   :800.0           Max.   :2.0000   Max.   :1.0000
 jazyk      vyska           vaha        pohlavi     ztrata          test
 A:43   Min.   :155.0   Min.   :48.00   M:34    Min.   :0.000   Min.   :34.00
 F: 3   1st Qu.:168.3   1st Qu.:59.00   Z:28    1st Qu.:1.000   1st Qu.:56.25
 N:13   Median :176.0   Median :65.00           Median :3.000   Median :71.00
 R: 3   Mean   :176.3   Mean   :70.50           Mean   :3.113   Mean   :68.43
        3rd Qu.:183.0   3rd Qu.:84.75           3rd Qu.:5.000   3rd Qu.:80.00
        Max.   :195.0   Max.   :99.00           Max.   :8.000   Max.   :97.00
```

Obrázek 9. Číselné charakteristiky pro všechny data

Statistics - Summaries - Numerical summaries

Tato funkce slouží pro výpočet průměru, směrodatné odchylky a námi zadaných kvantilů. Jak ukazuje obrázek číslo 10, tato funkce se dá použít pouze pro jedinou proměnnou. V okně se nám otevře nabídka jednotlivých charakteristik, které máme možnost spočítat. (průměr, směrodatnou odchylku nebo kvantil). V případě kvantilu se nám zde nabízí ještě možnost ručně dopsat kolika procentní kvantil potřebujeme. Obrázek číslo 11 znázorňuje výpočty pro proměnnou váha.



Obrázek 10. Vyplnení nabídky Numerical Statistics pro proměnnou váha

```
> mean(data$vaha, na.rm=TRUE)
[1] 70.5

> sd(data$vaha, na.rm=TRUE)
[1] 13.90807

> quantile(data$vaha, c( 0,.25,.5,.75,1 ), na.rm=TRUE)
   0%   25%   50%   75%  100%
48.00 59.00 65.00 84.75 99.00
```

Obrázek 11. Číselné charakteristiky pro proměnnou váha

Pokud bychom chtěli spočítat například průměrnou výšku podle pohlaví. Vybereme proměnnou „vyska“, klikneme na *Summarize by groups* a označíme proměnnou „pohlavi“. Na výstupu se nám vypočtou zadané charakteristiky zvlášť pro muže a zvlášť pro ženy. (na obrázku číslo 12 vidíme výpočet průměru a směrodatné odchylky)

```
> by(data$vyska, data$pohlavi, mean, na.rm=TRUE)
INDICES: M
[1] 182.7059
------------------------------------------------------------
INDICES: Z
[1] 168.5

> by(data$vyska, data$pohlavi, sd, na.rm=TRUE)
INDICES: M
[1] 6.032887
------------------------------------------------------------
INDICES: Z
[1] 6.647194
```

Obrázek 12. Průměr a směrodatná odchylka podle proměnné pohlaví

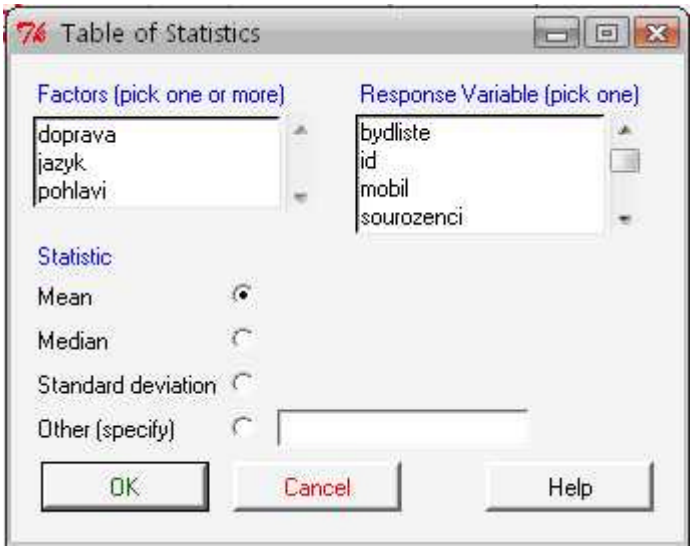
Statistics - Summaries - Frequency distribution

Vytvoří tabulku absolutních a relativních četností pro zvolené proměnné.

Statistics - Summaries - Table of statistics

Tato funkce je velmi užitečná, pokud chceme spočítat průměr, medián, směrodatnou odchylku či jinou statistickou veličinu zvlášť pro skupiny určené kategoriální proměnnou.

V kolonce **Factors** vybereme příslušnou kategoriální proměnnou a v kolonce **Response Variable** odpovídající numerickou proměnnou. V druhé části okna pak už jen označíme, kterou charakteristiku požadujeme vypočítat (viz.Obr.13).



Obrázek 13. Vyplnění nabídky Table of Statistics

Na obrázku číslo 14 vidíme výsledky průměrného počtu dosažených bodů v testu zvlášť pro muže a pro ženy.

```
> tapply(data$test, list(pohlavi=data$pohlavi), mean, na.rm=TRUE)
pohlavi
       M        Z
71.61765 64.55357
```

Obrázek 14. Průměrný počet dosažených bodů v testu pro proměnnou pohlaví

Statistics – Summaries – Correlation matrix

Pomocí tohoto příkazu můžeme zhotovit korelační matici.

Obrázek číslo 15 znázorňuje korelaci mezi proměnnými „vaha“ a „vyska“. Výsledky dle očekávaní vypovídají o poměrně těsném vztahu mezi výškou a váhou respondenta.

```
> cor(data[,c("vaha","vyska")], use="complete.obs")
           vaha     vyska
vaha  1.0000000 0.8856494
vyska 0.8856494 1.0000000
```

Obrázek 15. Korelace mezi proměnnými váha a výška

6.2. Kontingenční tabulky

Statistics - Contingency Tables - Two-way table

Vytvoří kontingenční tabulku ze dvou námi určených kategoriálních proměnných. V kolonce **row variable** označíme proměnnou, kterou chceme mít zobrazenou v řádcích a v kolonce **Column variable** proměnnou, kterou chceme mít ve sloupcích. Dále nám tabulka nabízí, zda chceme zobrazit procentuální vyjádření pro sloupce nebo pro řádky anebo jestli chceme zobrazit pouze tabulku četností (možnost **no percentages**).

V okně se nám rovněž nabízí možnost rovnou prozkoumat závislost či nezávislost námi vybraných kategoriálních veličin. K tomuto účelu slouží chí-kvadrát test nezávislosti (**Chi-square test of independence**).

Dále se nám zde nabízí možnost vytvořit tabulku očekávaných četností (**Print expected frequencies**).

Na ukázku si zkusíme zhotovit kontingenční tabulku pro proměnné jazyk a doprava. Výsledek vidíme na obrázku číslo 16.

```
> .Table <- xtabs(~jazyk+doprava, data=data)

> .Table
     doprava
jazyk  A  B  M  T  V
    A  1 10  8  6 18
    F  0  3  0  0  0
    N  0  5  1  1  6
    R  1  1  0  0  1
```

Obrázek 16. Kontingenční tabulka pro proměnné jazyk a doprava

Statistics - Contingency Tables - Multi-way table

Tato funkce nám umožní vytvořit několik kontingenčních tabulek ze 3 kategoriálních proměnných najednou.

Obrázek číslo 17 znázorňuje kontingenční tabulku pro proměnné jazyk, doprava a pohlaví.

```
> .Table <- xtabs(~jazyk+doprava+pohlavi, data=data)

> .Table
, , pohlavi = M

     doprava
jazyk  A  B  M  T  V
    A  1  5  4  2 12
    F  0  3  0  0  0
    N  0  2  1  0  3
    R  1  0  0  0  0

, , pohlavi = Z

     doprava
jazyk  A  B  M  T  V
    A  0  5  4  4  6
    F  0  0  0  0  0
    N  0  3  0  1  3
    R  0  1  0  0  1
```

Obrázek 17. Kontingenční tabulky pro proměnné jazyk, doprava a pohlaví

Statistics - Contingency Tables - Enter and analyze two-way table

Program nám nabízí i možnost vytvořit si vlastní kontingenční tabulku.V nabídce si můžeme zvolit počet sloupců a řádků (od 1 do 10) a pak už stačí jen zapsat všechny hodnoty do připravené tabulky. Tabulku pak můžeme nechat vytisknou buď procentuálním vyjádřením pro sloupce nebo řádky anebo ponechat četnosti přesně tak,

jak jsme je zapsali do tabulky. Samozřejmě i zde si rovnou můžeme otestovat závislost či nezávislost námi zadaných kategoriálních veličin pomocí chí-kvadrát testu nezávislosti.

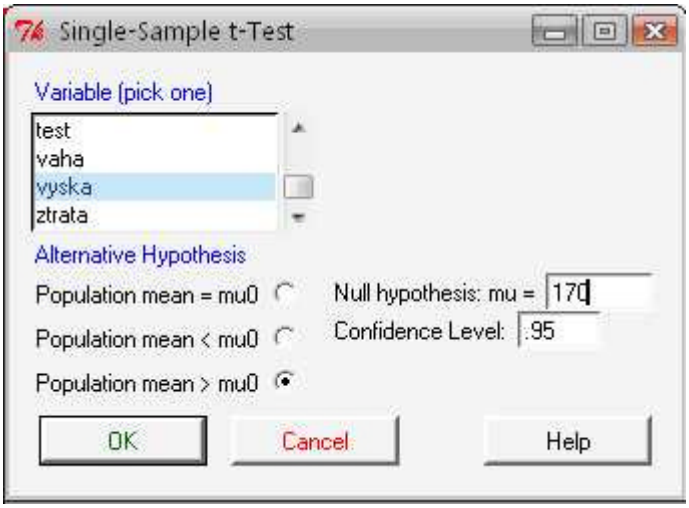
6.3. Testy o střední hodnotě

6.3.1. Jednovýběrový t-test

Statistics - Means - Single-sample t-test

Nejedná se o nic jiného než o nám dobře známý jednovýběrový t-test. V kolonce **Variable** vybereme jednu proměnnou, kterou chceme testovat. Poté vyplníme nulovou a alternativní hypotézu. A pokud je potřeba, můžeme v kolonce **Confidence Level** změnit hladinu spolehlivosti pro náš test.

Na obrázku číslo 18 vidíme, jak bychom vyplnili tabulku, pokud bychom dostali následující úkol: Na 95 procentní hladině spolehlivosti otestujte, zda je výška námi oslovených respondentů v průměru vyšší než 170 centimetrů.



Obrázek 18. Vyplnění nabídky pro jednovýběrový t-test

Výsledek testu si můžeme prohlédnout na obrázku číslo 19. P hodnota nám vyšla menší než 5 procent, tudíž zamítáme nulovou hypotézu ve prospěch alternativní hypotézy. S 95 procentní spolehlivostí lze tvrdit, že výška oslovených respondentů je v průměru vyšší než 170 cm.

```
> t.test(data$vyska, alternative='greater', mu=170, conf.level=.95)

        One Sample t-test

data:  data$vyska
t = 5.2196, df = 61, p-value = 1.139e-06
alternative hypothesis: true mean is greater than 170
95 percent confidence interval:
 174.2775      Inf
sample estimates:
mean of x
 176.2903
```

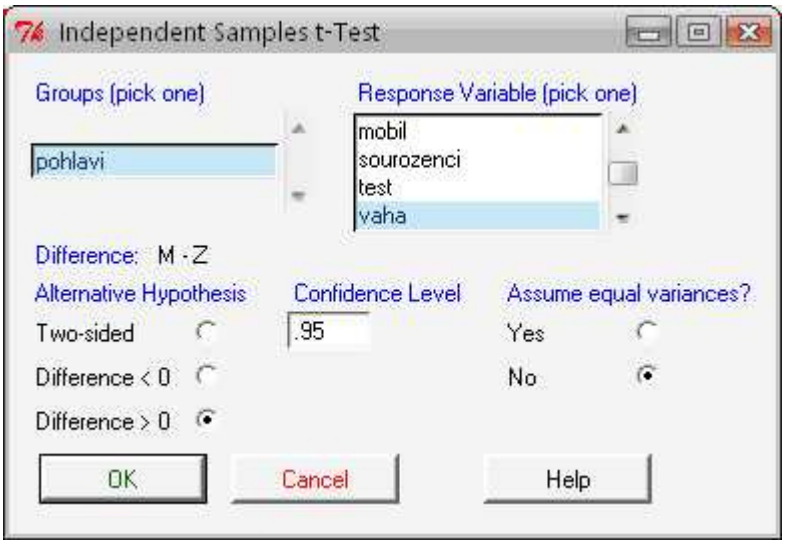
Obrázek 19. Výsledek jednovýběrového t-testu pro proměnnou výška

6.3.2. Dvouvýběrový t-test

Statistics - Means - Independent-samples t-test

Tento test reprezentuje dvouvýběrový t-test nebo-li test o shodě průměrů pro dva nezávislé výběry. Před tímto testem je vždy nejprve zapotřebí otestovat, jestli mají oba vzorky stejné rozptyly. K testování rozptylů použijeme F-test (bližší informace o tomto testu si povíme později.). Podle výsledků F-testu pak budeme schopni správně vyplnit kolonku **Assume equal variance?**. Ostatní kolonky vyplníme podobně jako u jednovýběrového t-testu. V kolonce **Groups** vyplníme, podle které proměnné budeme testovat (musí obsahovat pouze 2 různé možnosti) a v kolonce **Response Variable** označíme proměnnou, kterou chceme testovat. Jak postupovat při správném vyplnění tabulky si nyní ukážeme na následujícím příkladě.

Na 95 procentní hladině spolehlivosti otestujte, zda mají muži v průměru vyšší váhu než ženy. Budeme předpokládat, že rozptyly obou vzorků nelze považovat za shodné, proto si zaškrtneme v kolonce **Assume equal variance?** možnost **no**. Testujeme váhu podle pohlaví, tudíž v kolonce **Groups** označíme pohlaví a v kolonce **Response Variable** použijeme proměnnou váha. V tomto testu si musíme dát pozor na správné zaškrtnutí alternativní hypotézy. Počítač nám totiž řadí jednotlivé skupiny, podle kterých testujeme, v abecedním pořadí. Pokud bychom tedy chtěli například otestovat, jestli ženy mají vyšší hmotnost než muži, museli bychom otočit nerovnost a jako alternativní hypotézu zaškrtnout druhou možnost (diference< 0, tzn. muži mají nižší hmotnost než ženy). Pro náš případ nic takového řešit nemusíme, vyplníme tedy tabulku přesně tak, jak můžeme vidět na obrázku číslo 20.



Obrázek 20. Vyplnění nabídky pro dvouvýběrový t-test

Výsledky testu si můžeme prohlédnout na obrázku číslo 21. P hodnota je menší než 5 procent, tudíž opět zamítáme nulovou hypotézu ve prospěch alternativní. S 95 procentní pravděpodobností můžeme tvrdit, že muži mají v průměru vyšší váhu než ženy.

```
> t.test(vaha~pohlavi, alternative='greater', conf.level=.95, var.equal=FALSE, data=data)

        Welch Two Sample t-test

data:  vaha by pohlavi
t = 11.0422, df = 48.721, p-value = 3.655e-15
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 18.83573      Inf
sample estimates:
mean in group M mean in group Z
       80.52941        58.32143
```

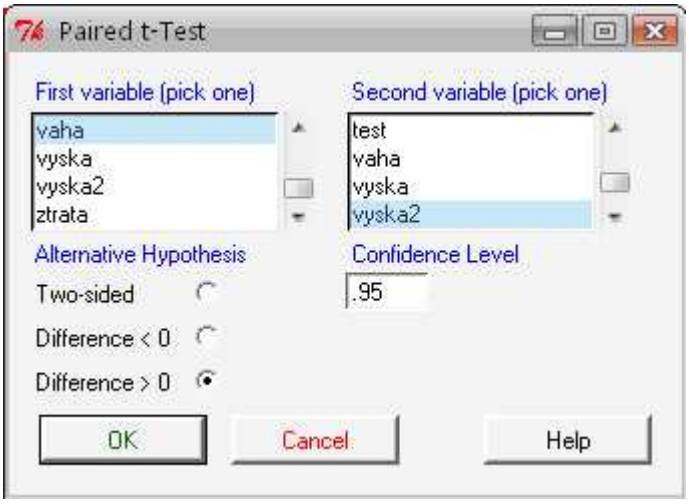
Obrázek 21. Výsledek dvouvýběrového t-testu

6.3.3. Párový t-test

Statistics – Means – Paired t-test

Tento test představuje párový t-test neboli test o shodě průměrů pro dva závislé výběry. Na příkladě si zkusíme otestovat, zda je výška zmenšená o 105 srovnatelná s váhou studenta. Nejprve si pomocí příkazu **Compute new variable** (najdeme v nabídce data - manage variables in active data set) vytvoříme novou proměnou „vyska2“ (=vyska-105). Nyní otestujeme na 95 procentní hladině významnosti, zda je váha respondentů

v průměru vyšší než námi vytvořené pravidlo pro proměnnou výška2. Vyplnění tabulky je velice snadné, stačí pouze označit proměnné, které chceme testovat, hladinu spolehlivosti a alternativní hypotézu (viz Obr.22).



Obrázek 22. Vyplnění nabídky pro Párový t-test

```
> t.test(data$vaha, data$vyska2, alternative='greater', conf.level=.95, paired=TRUE)

        Paired t-test

data:  data$vaha and data$vyska2
t = -0.8826, df = 61, p-value = 0.8096
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -2.285859       Inf
sample estimates:
mean of the differences
             -0.7903226
```

Obrázek 23. Výsledek párového t-testu

Ve výsledku je p hodnota vyšší než 5 procent, nulovou hypotézu tudíž nezamítáme. S 95 procentní spolehlivostí jsme neprokázali, že váha respondentů je vyšší než námi vytvořené nové pravidlo pro proměnnou výška.

6.3.4. ANOVA jednoduchého třídění

Statistics - Means - One-way ANOVA

Díky této funkci můžeme otestovat závislost numerické proměnné na kategoriální proměnné. Z našich dat jsme se rozhodli zjistit závislost výsledků testů na volbě jazyka a výsledek je znázorněn níže. Obr.24., který ukazuje, že výsledek testu na volbě jazyka nezávisí a to díky hodnotě Pr(>F), která je větší než 5%.

```
Analysis of Variance Table

Response: test
           Df  Sum Sq Mean Sq F value Pr(>F)
jazyk       3   879.9   293.3  1.1096 0.3526
Residuals 58 15332.0   264.3

> tapply(data$test, data$jazyk, mean, na.rm=TRUE) # means
       A        F        N        R
70.29070 72.00000 61.15385 69.66667

> tapply(data$test, data$jazyk, sd, na.rm=TRUE) # std. deviations
       A        F        N        R
15.35734 21.37756 18.75654 12.05543

> tapply(data$test, data$jazyk, function(x) sum(!is.na(x))) # counts
 A  F  N  R
43  3 13  3
```

Obrázek 24. Výsledek pro ANOVu jednoduché třídění

6.3.5. Vícerozměrná ANOVA

Statistics - Means - Multi-way ANOVA

Pokud potřebujeme zjistit závislost numerické proměnné na více proměnných kategoriálních, musíme zvolit ANOVU vícerozměrnou. Po otevření dialogového okna si budeme moci v levé části tabulky pomocí klávesnice „CTRL“ a kliknutí vybrat námi požadované proměnné a výběr potvrdit. Výstup nám opět ukazuje průměry, směrodatné odchylky a počet výskytu v souboru viz Obr.25. Podle hodnot Pr(>F) vidíme, že ani v tomto případě se nám nepotvrdila závislost výsledku testu na vlastnictví mobilu či pohlaví, i když u druhé zmiňované by se výsledek blížil hranici 5% mnohem výrazněji.

```
> Anova(lm(test ~ mobil*pohlavi, data=data))
Anova Table (Type II tests)

Response: test
              Sum Sq Df F value  Pr(>F)
mobil         256.5  1  0.9878 0.32441
pohlavi       916.1  1  3.5283 0.06536 .
mobil:pohlavi 129.9  1  0.5002 0.48225
Residuals     15059.4 58
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> tapply(data$test, list(mobil=data$mobil, pohlavi=data$pohlavi), mean, na.rm=TRUE) #
means
       pohlavi
mobil         M        Z
  ma   72.40909 66.19565
  nema 70.16667 57.00000

> tapply(data$test, list(mobil=data$mobil, pohlavi=data$pohlavi), sd, na.rm=TRUE) # std.
deviations
       pohlavi
mobil         M        Z
  ma   13.76904 17.68714
  nema 18.14003 12.00000

> tapply(data$test, list(mobil=data$mobil, pohlavi=data$pohlavi), function(x) sum(!is.na(x)))
# counts
       pohlavi
mobil   M  Z
  ma   22 23
  nema 12  5
```

Obrázek 25. Výsledek pro vícerozměrnou ANOVu

6.4. Testy o proporci

Statistics – Proportions- Single-sample proportion test

Test o proporci neboli proporcionální test.

Zde si na 95 procentní hladině spolehlivosti otestujeme, zda muži tvoří nadpoloviční většinu z celkového počtu všech respondentů.

Tabulku vyplníme klasickým způsobem, jak jsme zvyklí z předchozích testů, jediné políčko, které stojí za zmínku, je v tomto případě nulová hypotéza. Nulová hypotéza zní, že muži na škole tvoří jednu polovinu všech žáků, tudíž do políčka vepíšeme 0,5 (pokud bychom změnili zadání a ptali se například na jednu čtvrtinu, do políčka bychom zaznamenali číslo 0,25).

Dále si u tohoto testu musíme dát pozor na správné pořadí výběrů, software nám zde automaticky předurčuje abecední pořadí (M, Z). Pokud označíme u alternativní hypotézy znaménko „<“, náš program příkaz automaticky zpracuje jako M<Z. Je třeba dávat si velký pozor na znění zadání.

Vraťme se nyní zpět k našemu testu, výsledky můžeme pozorovat na obrázku číslo 26.

```
> .Table <- xtabs(~ pohlavi , data= data )

> .Table
pohlavi
 M  Z
34 28

> prop.test(rbind(.Table), alternative='greater', p=.5, conf.level=.95, correct=FALSE)

        1-sample proportions test without continuity correction

data:  rbind(.Table), null probability 0.5
X-squared = 0.5806, df = 1, p-value = 0.2230
alternative hypothesis: true p is greater than 0.5
95 percent confidence interval:
 0.4445823 1.0000000
sample estimates:
        p
0.5483871
```

Obrázek 26. Výsledek testu o proporci

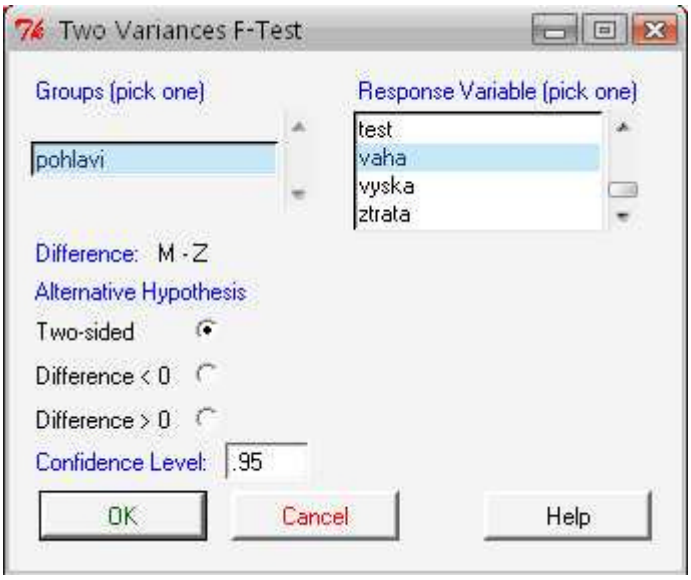
P hodnota je vyšší než 5 procent, tudíž nezamítáme nulovou hypotézu a s 95 procentní spolehlivostí se dá říci, že muži tvoří na škole zhruba jednu polovinu z celkového počtu všech žáků.

6.5. Testy o rozptylu

Statistics – Variance- Two-variances F-test

Testuje shodu rozptylů dvou výběrů. Tabulku vyplníme podobně jako u t-testů.

Nyní se můžeme vrátit zpět k příkladu, který byl uveden u dvouvýběrového t-testu. Potřebovali jsme otestovat, zda mají muži stejný rozptyl váhy jako ženy. Tabulku bychom vyplnili přesně jak vidíme na obrázku číslo 27.



Obrázek 27. Vyplnění tabulky pro test o Rozptylu

Výsledky testu na obrázku 28 prokazují, že náš předpoklad byl správný. Rozptyl obou výběrů nemůžeme považovat za shodný.

```
> var.test(vaha ~ pohlavi, alternative='two.sided', conf.level=.95, data=data)

        F test to compare two variances

data:  vaha by pohlavi
F = 4.5507, num df = 33, denom df = 27, p-value = 0.0001306
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 2.156213 9.347395
sample estimates:
ratio of variances
          4.550660
```

Obrázek 28. Výsledek testu o rozptylu

Statistics – Variance - Bartlett's test

Slouží podobně jako Leveneuv test ke zjišťování homoskedasticity. Podmínkou tohoto testu je však zachování normality dat.

Statistics – Variance - Levene's test

Pomocí tohoto testu zjišťujeme homoskedasticitu dat.

6.6. Neparametrické testy

Statistics – Nonparametric tests- Two-sample Wilcoxon test

Dvouvýběrový Wilcoxonuv test, který je neparametrickou obdobou dvouvýběrového t-testu, tzn. že nevyžaduje předpoklad normality.

Statistics – Nonparametric tests - Paired-samples Wilcoxon test

Podobně jako předchozí test, je i Párový Wilcoxonuv test neparametrickou obdobou párového t-testu.

Statistics – Nonparametric tests - Kruskal-Wallis test

Kruskaluv-Wallisuv test je neparametrickou obdobou ANOVy jednoduchého třídění.

Více informací ohledně těchto testů nalezneme v publikacích od autorů Komárka, Komárkové a Bíny [6], [8]. Uvedené materiály je možné stáhnout ze školního disku „K“ respektive z K:\PED\KMIH\verejny\kmm\Komarek\skriptum\.

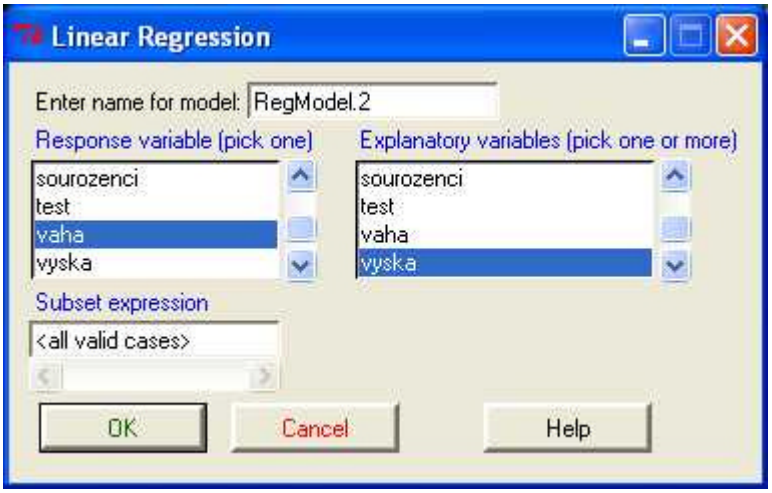
6.7. Vícerozměrná analýza

Vícerozměrnou analýzou se v tomto textu zabývat nebudeme. Bližší informace k této problematice můžete nalézt například v knihách od autora Hebáka . Viz [3], [4], [5].

6.8. Fitování modelu

Statistics –Fit models - Linear regression

Jako vhodný příklad Lineární regrese můžeme použít závislost váhy studentů na jejich výšce. Zadání proměnných vidíme na Obr.29 a následný výstup zobrazuje Obr.30.



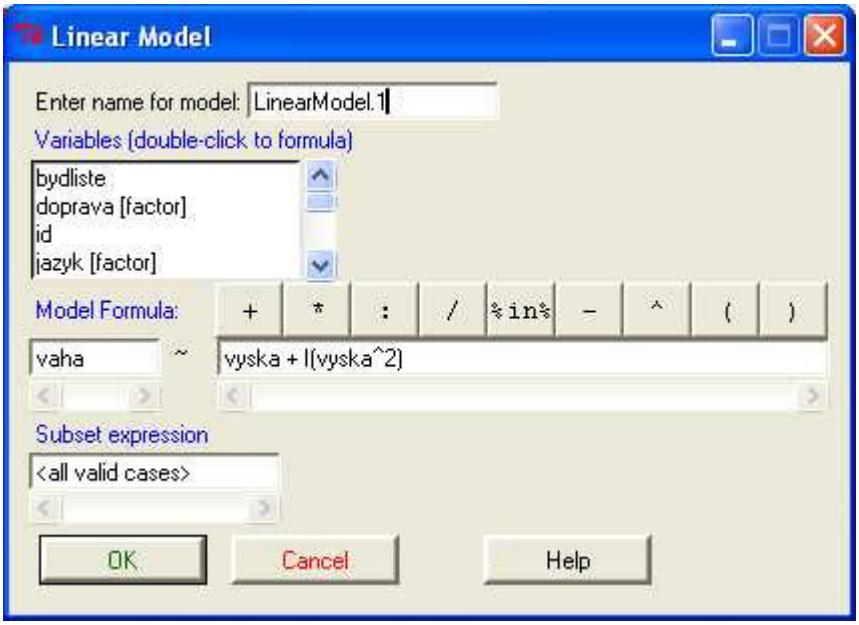
Obrázek 29. Vyplnění nabídky pro lineární regresi

```
Call:
lm(formula = vaha ~ vyska, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-14.1038  -4.8635   0.7614   4.2087  16.2421

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -158.33701   15.51160  -10.21 9.76e-15 ***
vyska          1.29807    0.08786   14.77  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.512 on 60 degrees of freedom
Multiple R-Squared: 0.7844,     Adjusted R-squared: 0.7808
F-statistic: 218.3 on 1 and 60 DF,  p-value: < 2.2e-16
```

Obrázek 30. Výsledek pro lineární regresi

Statistics –Fit models - Linear model

I v případě lineárního modelu můžeme postup ukázat na závislosti stejných proměnných, jako tomu bylo v případě lineární regrese, přičemž bude výstup naprosto totožný. Hlavní rozdíl ale uvidíme v dialogovém okně, které se nám zobrazí ihned po zadání postupné cesty uvedené v nadpise. Zmíněné zobrazení je znázorněno na Obr.31.



Obrázek 31. Vyplnění nabídky pro lineární model

Tabulka je rozšířena o část s názvem *Model formula,* která nám poslouží jako příkazový řádek pro tvorbu funkčního předpisu. Tento předpis se dá využít například pokud budeme chtít zjistit, zda jedna proměnná závisí na té druhé kvadraticky. Výstup ukazuje Obrázek 31.

```
Call:
lm(formula = vaha ~ vyska + I(vyska^2), data = Dataset)

Residuals:
     Min       1Q   Median       3Q      Max
-14.3540  -4.5423  -0.4267   3.8900  15.3210

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 365.656472 248.382574   1.472   0.1463
vyska        -4.679494   2.829537  -1.654   0.1035
I(vyska^2)    0.016999   0.008043   2.114   0.0388 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.332 on 59 degrees of freedom
Multiple R-Squared: 0.7996,     Adjusted R-squared: 0.7928
F-statistic: 117.7 on 2 and 59 DF,  p-value: < 2.2e-16
```

Obrázek 32. Výsledek pro lineární model

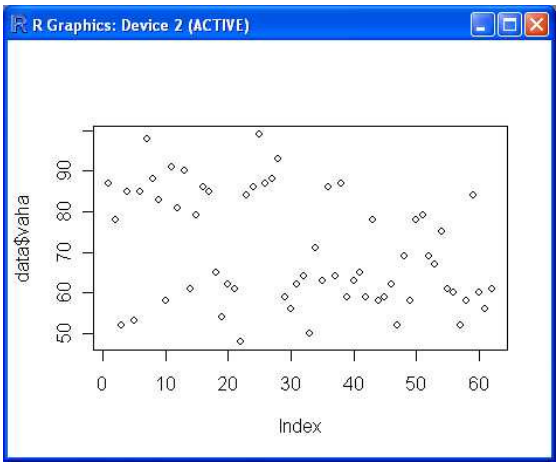
Do příslušného políčka dostaneme požadované proměnné postupným „dvojklikáním“ na samotné proměnné, přičemž se nám do vzorce bude automaticky vkládat znaménko „+“. Pokud bude však chtít vzorec jiný, než pouhý součet, máme ostatní znaky k dispozici v nabídce. Výsledky tohoto konkrétního příkladu vidíme na Obrázku 32, ze kterého je patrné, že váha opravdu na výšce kvadraticky závisí, jelikož je hodnota Pr(>|t|) menší než 5%.

7. Grafy

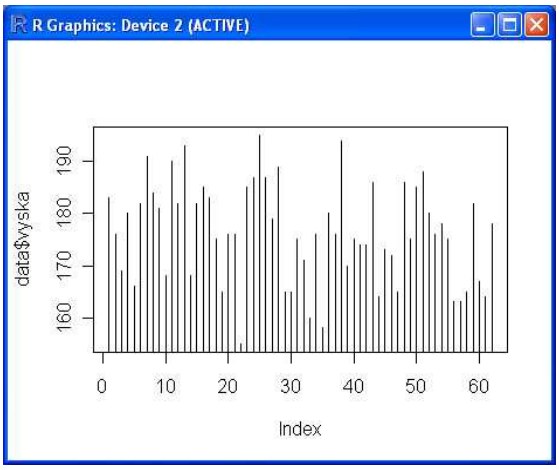
Další velice užitečnou funkcí, kterou můžeme snadno využít, je zobrazování dat pomocí grafů. V mnohých případech si pak lépe uvědomíme strukturu dat, než by tomu bylo pouze v „tabulkové“ podobě. Proces zadání příkazu a výběr požadované proměnné pro zobrazení grafu je opět velice jednoduchý a rychlý. Máme zde na výběr celou škálu různých typů grafu, takže si budeme moci vždy vybrat ten nejpřehlednější pro nás. Ukážeme si zde alespoň nějaké typické zástupce grafů, se kterými se pravděpodobně budeme při práci setkávat nejčastěji. Na úvod této kapitoly bychom si ještě měli zmínit jednu poznámku, a sice, kde se nám budou vytvořené grafy zobrazovat. Ať už si vybereme jakýkoliv druh grafu, tak po zadání proměnných a jejich následném potvrzení se na první pohled nic nestane. To je však naprosto v pořádku a my se nemusíme nijak znepokojovat faktem, že námi požadovaná akce se nezdařila. Veškeré grafické výstupy se nám totiž zobrazují do původního okna programu R s názvem „RGui“, do kterého se jednoduše přepneme pomocí lišty ve Windows.

7.1. Index plot

Pro vykreslení načtených dat v podobě grafu nám slouží hned první příkaz v záložce *Graphs* tj. *Index plot.* Kromě výběru požadované proměnné si zde můžeme vybrat ze dvou druhů typu zobrazení, a to zatrhnutím možnosti *Spikes* nebo *Points.* Jak už je z názvu patrné, první typ zobrazuje hodnoty v podobě úseček, oproti možnosti druhé, značící pouze body(viz Obr.33 a 34).



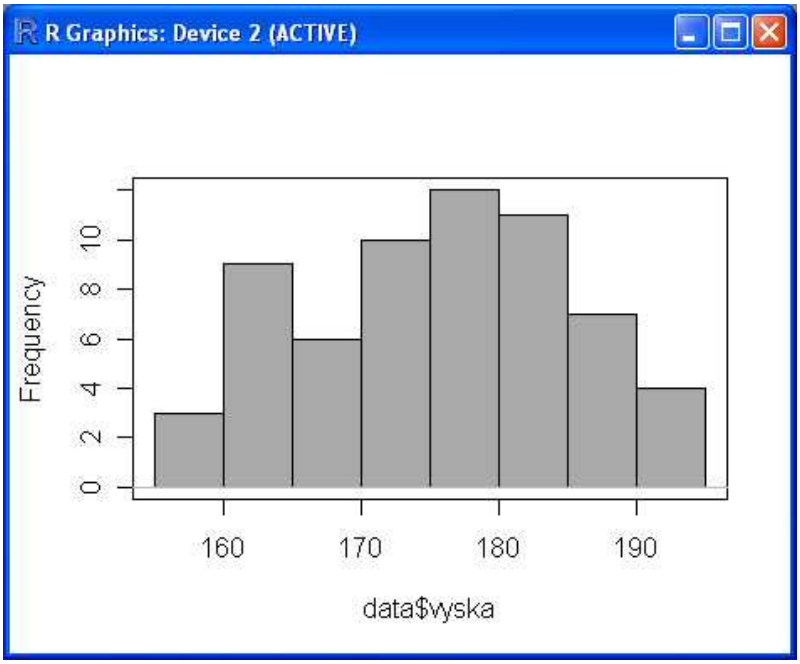
Obrázek 33. Index plot - body



Obrázek 34. Index plot - úsečky

7.2. Histogram

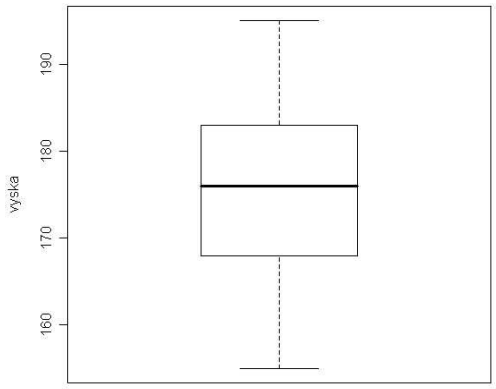
Dalším velice často používaných zobrazením je *Histogram*. U tohoto grafu si opět můžeme volit mezi několika druhy. Stejnou formou jako u grafu předchozího zaškrtneme tu variantu, kterou požadujeme. Na výběr máme zobrazení dle počtu výskytu v datovém souboru, procentuálně a nebo pomocí hustoty. Pro znázornění jsme vybrali zobrazení dle počtu a proměnnou byla výška studentů. Výstup je znázorněn na Obr.35.



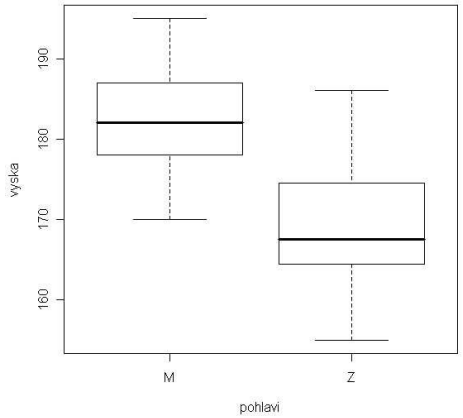
Obrázek 35. Histogram

7.3. Krabičkový graf

Boxplot poskytuje základní informace o analyzovaných datech. V tabulce stačí pouze označit numerickou proměnnou, ze které si přejeme graf vytvořit. Po kliknutí na **plot by groups** můžeme vytvořit grafy pro skupiny (viz. Obr.37 – grafy pro výšku mužů a žen), Horní a spodní hranice představují maximum a minimum, silná čára uprostřed označuje medián a „krabice“ kvartily.



Obrázek 36. Krabičkový graf

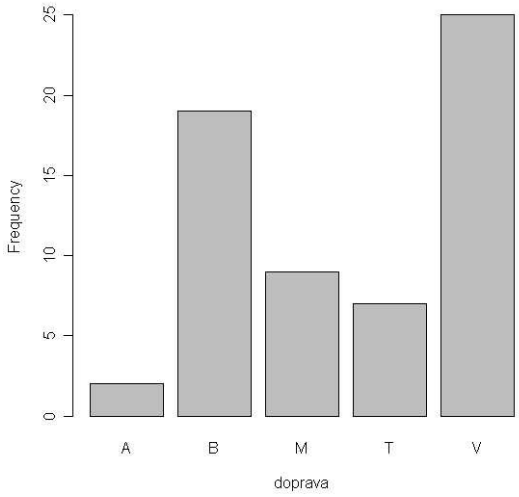


Obrázek 37. Krabičkový graf zvlášť pro muže a ženy

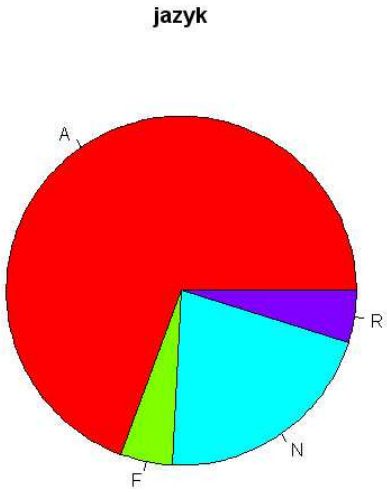
7.4. Koláčový a sloupcový graf

K zobrazení kategoriální proměnné můžeme využít dva druhy grafů. Buď koláčový graf (*Pie Chart*) nebo sloupcový graf(*Bar graph*). Pomocí grafu koláčového máme data rozdělena do kruhu, kde četnosti výskytu odpovídá určitá výseč, kdežto graf sloupcový data rozdělí do obdélníků, jejichž výška odpovídá absolutní četnosti. Při vytváření těchto typů grafů nám postačí pouze zadat proměnnou a o zbytek se již postará software

sám. Obrázky číslo 38 a 39 ukazují sloupcový graf proměnné doprava respektive graf koláčový pro proměnnou jazyk.



Obrázek 38. Sloupcový graf

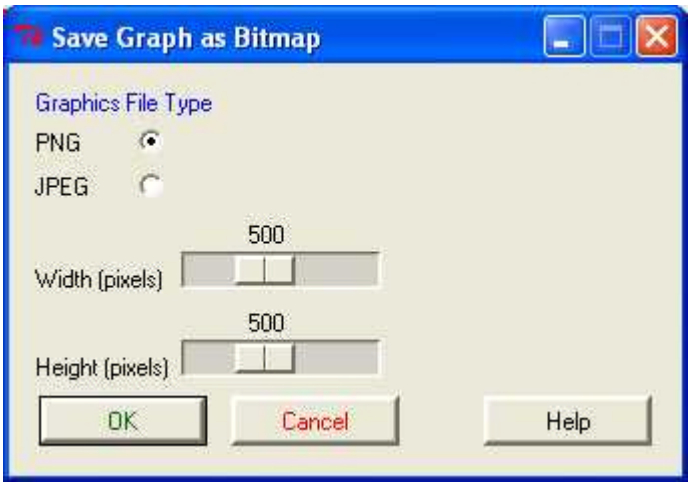


Obrázek 39. Koláčový graf

7.5. Ukládání grafů

Jako poslední, co zde zmíníme ohledně grafů, není funkce, která by svou užitečností na posledním místě být měla, nicméně pro přehlednost dodržíme pořadí příkazu tak, jak je tomu v menu R-commanderu. Tato část bude o ukládání grafů. Při naší práci je důležité veškerá rozhodnutí nebo výsledky podložit důkazy a grafy mohou být právě tím rozhodujícím důkazem. Z toho důvody je vhodné si grafy nějakým způsobem zálohovat, například pro pozdější kontrolu. Mohli bychom si samozřejmě grafy překreslit, nicméně

nikdy nebudou takto vytvořené materiály tak přesné jako originály a v dnešní době je takový způsob poněkud „staromódní“. Není tedy divu, že i tento problém je v dané nadstavbě vyřešen na úrovni. Pomocí příkazu *Save graph to file* můžeme naše grafické výsledky bez větších problém uložit. I tady máme několik možností, ze kterých si můžeme vybrat typ výstupního souboru. Zvolíme-li si *Save graph to file as bitmap..,* bude naším výstupem obrázek ve formátu jpeg či png dle naší následné volby viz Obr.40. Můžeme si zde nastavit i požadovanou velikost výsledného grafu v pixelech.



Obrázek 40. Uložení grafu

Při ukládání grafů je tvůrci myšleno i na příznivce formátu pdf, do kterého se nám výsledný obrázek exportuje kliknutím na příkaz *graph to file as PDF/Postscript/EPS.*.
I při výběru této možnosti nebudeme ochuzeni o volbu velikosti výsledného grafu.

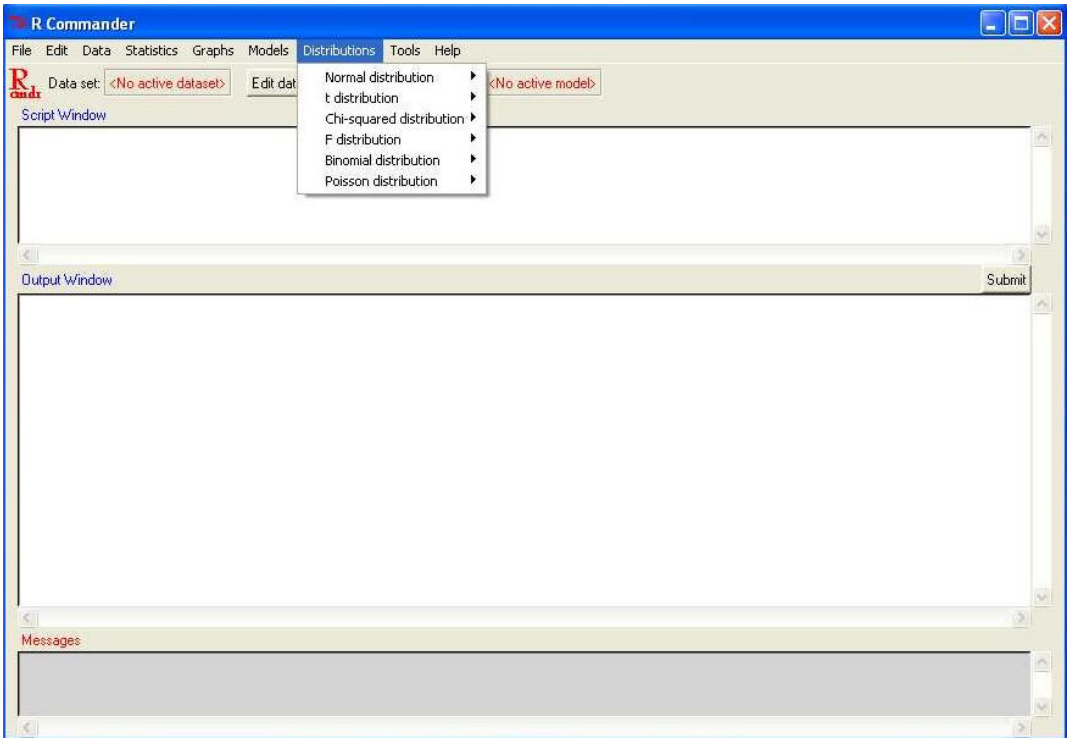
8. Modely

Jako další v pořadí v menu nalezneme záložku s názvem Models. V této sekci máme možnost pracovat s modely, upravovat je, testovat či jinak zjišťovat jejich vlastnosti. Jsou zde k dispozici jiné diagnózy, testy nebo grafy než jsme si doposud uváděli, ale jelikož jsou nad rámec našich potřeb při výuce na naší fakultě, nebudeme se o nich v této publikaci nijak více zmiňovat.

9. Pravděpodobnostní rozdělení

Další oblast statistických počtů, které nám pomáhá nadstavba R-commander spočítat patří i náhodné veličiny. Bez nutnosti znalosti různých složitých příkazů můžeme snadno a rychle zjistit hodnoty základních druhů spojitých i diskrétních náhodných veličin. Z diskrétních veličin to je rozdělení Binomické a Poissonovo, ze spojitých potom Normální, chí-kvadrát rozdělení, Studentovo a Fischerovo rozdělení.

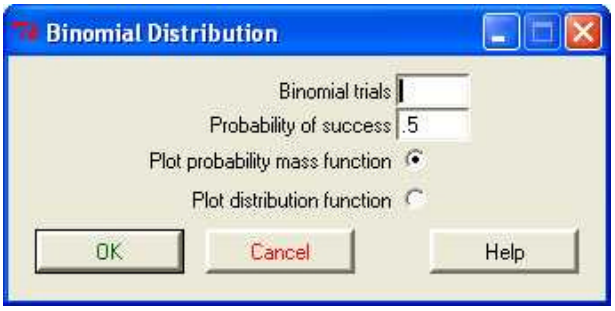
Kliknutím na nápis *Distributions* v hlavním menu se nám objeví nabídka, ve které si budeme moci vybrat druh rozdělení dle naší potřeby jak je vidět na Obr. 41.



Obrázek 41. Pravděpodobnostní rozdělení

U většiny rozdělení dále máme možnost volit mezi výpočtem distribuční (probability), kvantilové (quantile) a pravděpodobnostní (density) funkce, respektive hustotu pravděpodobnosti. Podle zvoleného rozdělení doplníme příslušné potřebné hodnoty jakými mohou být kromě počtů pokusů, výrobků či odpovědí (Variable Values) střední hodnota (mean), rozptyl (standart deviation), stupeň volnosti (degrees of freedom), pravděpodobnost úspěchu (probability of success) nebo počet příznivých pokusů (trials). Funkce *Lower* respektive *Upper tail* dole pod tabulkou nám umožňuje vypočítat doplňkový jev.

Kromě výpočtů zde máme k dispozici také vykreslení grafu funkce. Dialogové okno, do kterého zadáme potřebné údaje, vyvoláme potvrzením tlačítka *Plot* u příslušného rozdělení viz Obr.42. Zde si můžeme zvolit i druh grafu. Kromě zobrazení pravděpodobností jednotlivých možností (Plot probability mass function) můžeme nechat sestrojit také graf kumulativních součtů (Plot distribution function). Pro zobrazení grafu se budeme muset přepnout z R Commanderu do dialogového okna RGui.



Obrázek 42. Zobrazení brafu funkce

10. Uložení dat

Po ukončení naší práce v R commander si můžeme všechny naše výpočty a statistické analýzy uložit. Můžeme si zálohovat buď vše, co jsme vytvořili nebo pouze nějakou část naší práci. Tím máme na mysli, že si můžeme vybrat, jestli si uložíme pouze příkazy, které nám software automaticky „vygeneroval“ během zadávání příkazů a nebo zda nám postačí si uložit jen jednotlivé výstupy z námi provedených operací. Ať už se rozhodneme pro jakoukoliv variantu, uložení nebude nijak složité a i pro běžného uživatele dnešních programů intuitivní. Jako u většiny aplikací se totiž provádím pomocí hlavní nabídky přes záložku *File*. Pak už si jen zvolíme co chceme uložit a podle toho se rozhodneme pro možnost *Save skript as* nebo *Save output as.* Jak už názvy napovídají, první příkaz nám uloží pouze skripty, druhý zase výstupy. Poslední, k čemu budeme vyzvání bude místo, kam požadujeme náš „výtvor“ uložit. V případě bychom si například omylem R Commander chtěli vypnout bez předchozího uložení změn, automaticky se nás aplikace bude dotazovat, zda máme zájem script i output uložit. Budeme-li chtít příště pokračovat v rozdělané práci, v tom samém skriptu, zvolíme při spuštění R Commanderu příkaz *Open skript file* opět v záložce *File.*

11. Závěr

Cíl a smysl této práce byl od počátku z naší strany jednoznačný a sice sestrojit, pokud možno, jednoduchý, přehledný návod na obsluhu nadstavby statistického softwaru R. Jelikož se tento program na naší fakultě aktuálně používá při výuce statistiky, doufáme, že bude i díky naší snaze pronikání našich kolegů do tajů statistiky zase o kousek snazší. Pro názornost a lepší možnost pochopení jsme doplnili publikaci značným množstvím obrázků, které mnohdy řeknou více než sáhodlouhé věty. Z pohledu studenta je to jistě vítaný krok a věříme, že tento „manuál“ bude hojně užíván spokojenými spolustudenty nejen VŠE . Naší snahou, jak už sama anotace ukazuje, nebylo pouze sestavení a sepsání jakéhosi návodu na obsluhu této statistické aplikace, ale také porovnání kladů a záporů nadstavby oproti základní verzi. Myslíme si, že mezi největší přednosti, které nadstavba R-commader skýtá, je minimální znalost hesel pro zadávání příkazů k provedení potřebné statistické operace. Když se podíváme na výstup, který nám program vygeneruje, můžeme si všimnout, že zápisy jsou mnohdy hodně krkolomné a této práce jsme tudíž díky nadstavbě ušetřeny. Další velký dík patří jistě přehledně udělanému *Menu*, které je logicky seřazeno a někdy nám pomůže zvolit ten správný model či příkaz. Na druhé straně je zde, podle našeho názoru, i pár nedostatků. Některé z nich jsou podstatné více, jiné jsou, dle nás, spíše formalitou nebo otázkou zvyku. Běžnému uživateli možná občas bude scházet možnost potvrzení příkazu pomocí stisknutí klávesy *Enter* a bude „nucen“ použít tlačítko *submit* a také například zobrazování grafů nebo nápovědy v původní okně softwaru R, je trochu nepraktické. Tyto nedostatky jsou ale mnohonásobně vynahrazeny výše zmíněnými výhodami, které nám při práci naopak ušetří spoustu drahocenného času. Mimo tyto „formality“ zde nalezneme i vcelku závažné nedostatky, a to zvláště pro uživatele využívající složitější operace či ty, kteří rádi používají jiné než „defaultní“ nastavení programu. Máme tím na mysli především omezenost rozsahu Menu ve zpracovávané nadstavbě, což znamená, že některé statistické výpočty jinak, než přímým zápisem příkazu nelze provést. V Menu je zkrátka nenalezneme. Ohledně nadstandardního nastavení můžeme uvést jeden příklad pro názornost, a sice uspořádání výsledných grafů. Mnohdy nám totiž ušetří čas a práci porovnání výsledků v podobě grafů, a to je pak jistě užitečné si tyto grafy nechat vykreslit buď pod sebe a nebo vedle, aby porovnání bylo snadnější. Ani tuto operaci bez znalosti příkazu zapsaného přímo v aplikaci R Gui neprovedeme. Z celkového hlediska se ale domníváme, že pro rozsah

učiva a probíraných kapitol statistiky na naší fakultě, je nadstavba R-commander pro studenty přehlednější, praktičtější a snadnější na obsluhu, s minimálními nároky na nutnost učení se, jinak nezbytných, statistických zápisů.

12. Seznam obrázků

Literatura

[1] Bartošová Jitka, Základy statistiky pro manažery, Vysoká škola ekonomická v Praze , Nakladatelství Oeconomica, 1. vydání, 20.1.2006.

[2] Fox John, Journal of Statistical Software. Dostupné z <http://socserv.mcmaster.ca/jfox/Misc/Rcmdr/>. Poslední úprava 24.6.2007.

[3] Hebák a kol., Vícerozměrné statistické metody (1), Informatorium, 2007

[4] Hebák a kol., Vícerozměrné statistické metody (2), Informatorium, 2007

[5] Hebák a kol., Vícerozměrné statistické metody (3), Informatorium, 2007

[6] Komárek Arnošt, Komárková Lenka, Statistická analýza závislostí s příklady v R, Skriptum pro přednášku 6MI221, 7.2.2007

[7] Komárek Arnošt, Úvod do statistiky – Text doplňující přednášky Statistika A a B, 23. září 2005.

[8] Komárková Lenka, Komárek Arnošt, Bína Vladislav, Základy analýzy dat a statistického úsudku s příklady v R, Skriptum pro přednášku 6MI221, 29.11.2006