



EVROPSKÁ UNIE  
Evropské strukturální a investiční fondy  
Operační program Výzkum, vývoj a vzdělávání

**MŠMT**  
MINISTERSTVO ŠKOLSTVÍ,  
MLÁDEŽE A TĚLOVÝCHOVY

# **Základy práce v programu R skrze R commander**

**Petr Soukup**



**2020**

## Informace o autorovi:

Petr Soukup

Univerzita Karlova, Fakulta sociálních věd, Institut sociologických studií

soukup@fsv.cuni.cz

*„Tento výstup lze užít v souladu s licenčními podmínkami Creative Commons BY 4.0 International (<http://creativecommons.org/licenses/by/4.0/legalcode>).“*



EVROPSKÁ UNIE  
Evropské strukturální a investiční fondy  
Operační program Výzkum, vývoj a vzdělávání

**MŠMT**  
MINISTERSTVO ŠKOLSTVÍ,  
MLÁDEŽE A TĚLOVÝCHOVY

## Stručný úvod pro práce s R skrze R commander

### Obecně o R

R, česky někdy familiérně R-ko, je prostředí, které je připraveno pro matematické a statistické výpočty. Bylo vyvinuto jako nekomerční alternativa ke statistickému paketu S jeho jméno (R) se odvozuje od prvního písmene křestních jmen otců zakladatelů. Často se označuje též jako tzv. R project. Výhodou R je jednak jeho volná dostupnost (freeware), jednak velká šíře výpočetních procedur, R dále vyniká zejména v grafických zobrazeních dat. R je také zcela otevřeným systémem, tj. kdokoli, kdo umí programovat, může do R přispět tzv. balíčky. Základem pro práci s R je tzv. base modul (jádro, viz dále) a skrze balíčky se rozšiřují výpočetní a grafické možnosti R. Nevýhodou R zejména pro uživatele z oblasti sociálních věd je jeho specifický programovací jazyk a téměř výhradní absence nabídek. Pro užívání R se tedy uživatel musí tento jazyk naučit a používat jej. Dokumentace k R i jednotlivým balíčkům je zpravidla poměrně technicistní, což dále komplikuje užívání R v oblasti sociálních věd.

### Nadstavby v R a k R

Poměrně záhy po uvedení R pochopili statistici, že pro jeho větší rozšíření, bude potřebné zvýšit uživatelský komfort. Cesty, jak toho dosáhnout, jsou různé. Jednak vznikly speciální balíčky v R obsahující sady nabídek (nejznámější je zřejmě R commander, který si detailně představíme první den semináře), jednak vznikla doplňková prostředí, které kombinují nabídky a usnadňují práci s příkazy (nejznámější a jednoznačně nejužívanější je R Studio, které si detailně představíme druhý den semináře). Kromě toho se někteří vývojáři rozhodli užít výpočetní algoritmy z R, ale udělat zcela samostatné prostředí pro jeho ovládání a zcela opustit příkazový jazyk R. V době psaní tohoto textu chvíli jsou k dispozici zejména JASP, Jamovi a BlueSky Statistic. Vždy jde o freeware, který je možný volně instalovat a užívat, nicméně až na výjimky neumožňují tyto nadstavby k R generovat sady příkazů, které byly užity pro výpočty, případně tyto příkazy zjednodušují a nelze je tak užít v základním prostředí R.





## Instalace R

Nejdříve se naučíme instalovat R (base modul) a poté si na příkladu R commanderu ukážeme, jak nainstalujeme různé dílčí balíčky.

Samotné R je k dispozici v mnoha verzích přímo na webu projektu R: <https://cran.r-project.org/>. Zde je najdeme rozcestník dle operačního systému (Linux, Mac či Windows). Uživatelé Mac si vyberou soubory ke stažení dle verze operačního systému (verze 10.6 a vyšší jsou odlišeny od starších). Uživatelé Windows vyberou volbu Base (nabízena jako první na stránce R for Windows), zde jim bude nabídnuta nejaktuálnější verze, nyní 4.0.2).<sup>1</sup> Při instalaci (po stažení a spuštění instalovaného souboru) je musíme zvolit dle operačního systému, zda instalovat 32-bitová nebo 64-bitová verze.

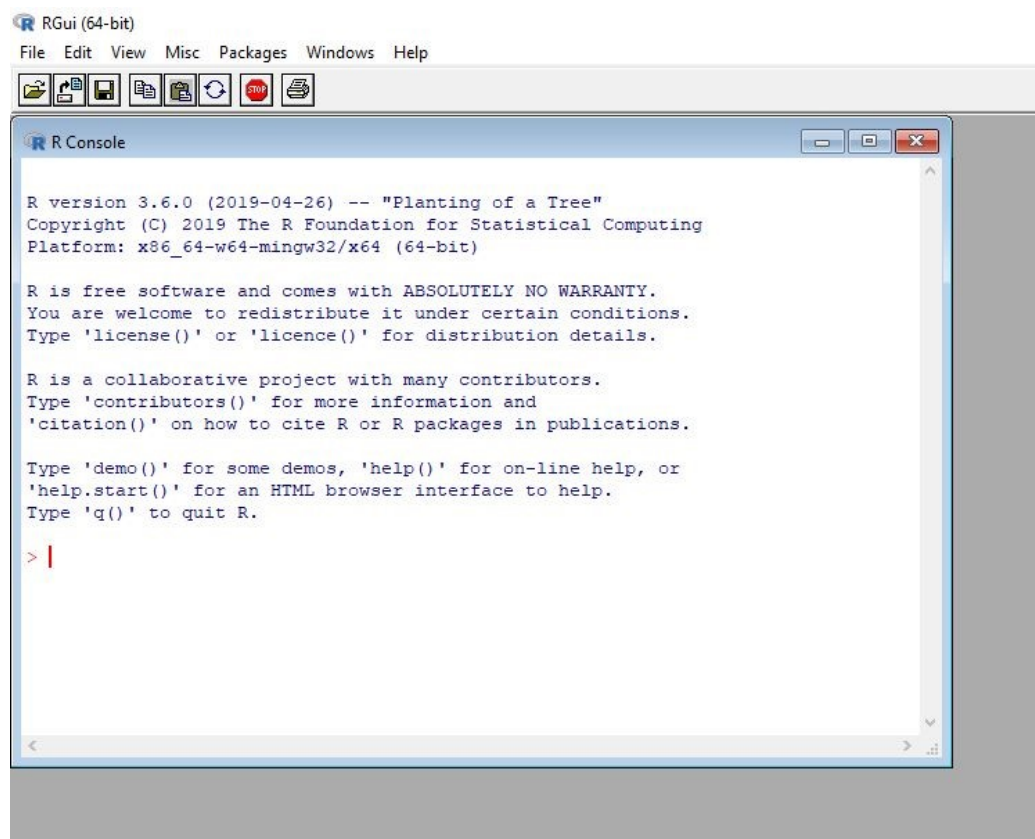
Po úspěšné instalaci se v programech objeví R (charakteristické ikonou s velkým modrým R).

Po kliknutí na ikonu se otevře základní prostředí R (viz obrázek 1).

---

<sup>1</sup> Jde o údaj na počátku října 2020.



**Obr. 1. Obrazovka R po prvním spuštění**

## Základní práce v R

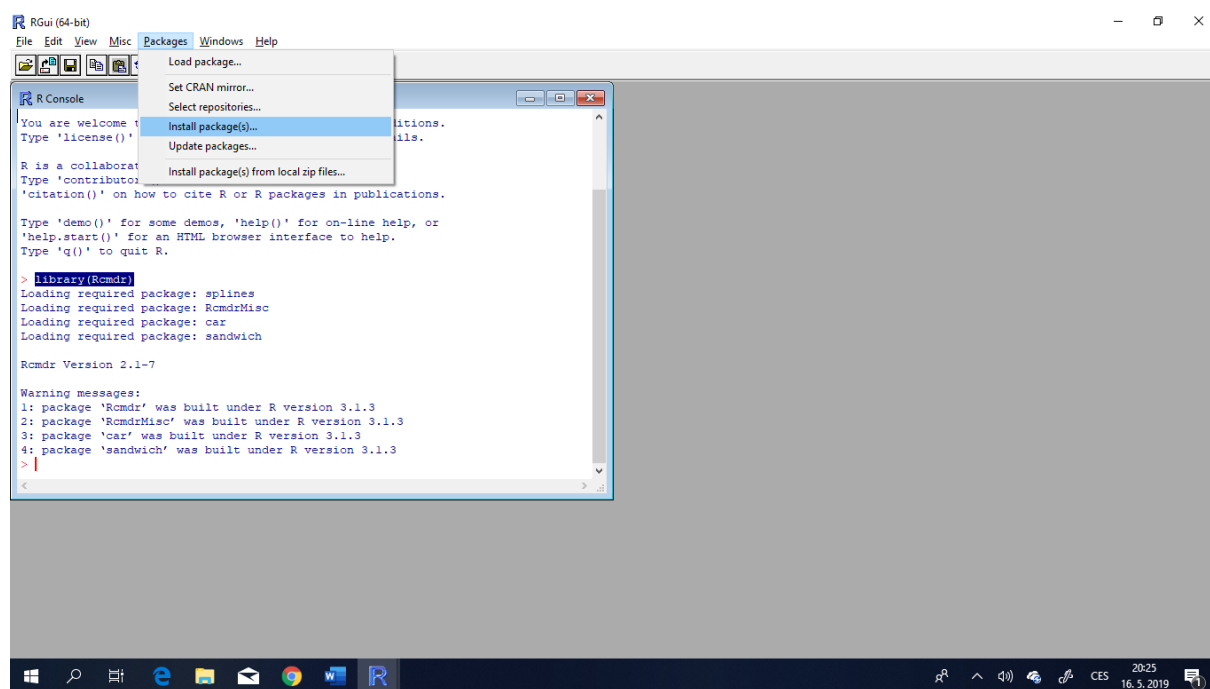
Práce v základním prostředí R vyžaduje, abychom do stavového řádku začali psát příkazy (typicky nejdříve pro načtení dat, dále pak pro jejich úpravy a analýzu). Skrze příkazy, ale též skrze nabídku lze provádět některá nastavení a zejména instalovat další balíčky. My jako první nainstalujeme R commander. Pro snazší postup ukážeme, jak provést instalaci skrze nabídky, posléze uvedeme i příkaz pro instalaci.

Z nabídky vybereme volbu *Packages*, která slouží instalaci jednotlivých balíčků. Instalaci lze provést buď online (přímo z webu R projektu) nebo offline (nejdříve stáhneme balíček ve formě zip souboru a poté instalujeme). Lepší je vždy při práci s R být online (pro možnost přidání dalších balíčků či jejich aktualizaci, proto si ukážeme tuto variantu (balíčky v R jsou poměrně malé soubory, nehrozí tedy stahování příliš velkého objemu dat). Před instalací je lepší si nastavit zrcadlo (tj. server), ze kterého se budou balíčky instalovat. V nabídce *Packages-Set Cran mirror* volíme zrcadlo ze země, kde se právě nacházíme (pro ČR jsou k



dispozici jedno až dvě zrcadla). Po volbě serveru nainstalujeme R commander, označený v R jako *Rcmdr*. Skrze nabídku *Packages-Install Package* nalezneme *Rcmdr* a spustíme instalaci. Poté musíme být zpravidla několik minut trpěliví (řádově minuty). Po dokončení instalace by měla obrazovka vypadat podobně jako na obrázku 2

**Obr. 2. Obrazovka po nainstalování Rcmdr do R**



Nainstalování balíčku lze ověřit v nabídce *Packages-Load Packages*, která obsahuje seznam všech úspěšně nainstalovaných balíčků. Pokud zde nalezneme *Rcmdr*, postupovali jsme správně.

Zde je vhodné upozornit, že balíčky stejně jako R se čas od času (zpravidla vícekrát ročně aktualizují). Stává se tak poměrně často, že pokud neupravíte novější verze samotného R (base modul), tak vám nepůjde některý balíček (jeho inovovaná verze instalovat). Řešením je užívat vždy jednu z nejnovějších verzí R. Není problém mít na jednom počítači více verzí R, nijak se mezi sebou nevadí, zabírají jen místo na pevném disku (R je ale i s balíčky poměrně nenáročný na místo na disku).

Pro úplnost ještě ukažme, jak bychom instalovali balíček *Rcmdr* příkazem:





```
install.packages("Rcmdr", dependencies=TRUE) 2
```

Obecná fráze `install.packages` sděluje programu, že chceme instalovat, v závorce v uvozovkách je pak uveden oficiální název balíčku v řeči R. Náš příkaz má ještě jednu část `dependencies=TRUE`, která znamená, že kromě balíčku *Rcmdr* se nainstalují i ostatní balíčky, které jsou pro R commander potřebné. Typické balíčky v R jsou stavebnicové konstrukce a pro své fungování potřebují další balíčky (někdy jich jsou i desítky), proto je instalace těchto doprovodných balíčků potřebná.

### **Spuštění R commanderu**

Poté co se balíček v R nainstaluje je třeba pro práci s ním jej ještě aktivovat. Tuto operaci lze opět provést skrze nabídku (*Packages-Load Package*), nebo příkazem.

Pro aktivaci R commanderu tedy můžeme napsat příkaz:

```
library(Rcmdr) 3
```

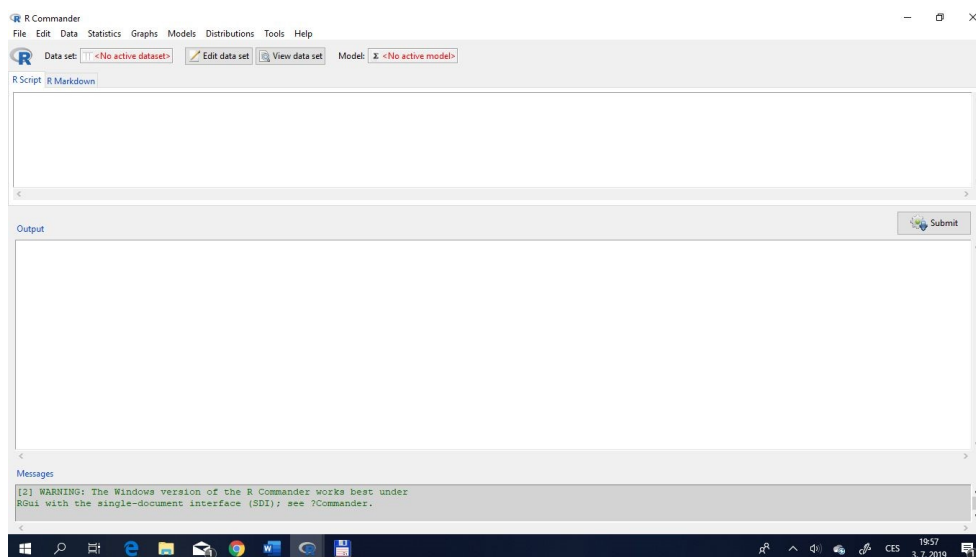
Po aktivaci balíčku R commander se otevře nové okno a v něm bude samotné prostředí R commanderu pro analýzu dat (obrázek 3).

---

<sup>2</sup> V R se pro příkazy, hlášky i výstupy používá font *Courier New*, proto jej i zde budeme k těmto účelům užívat.

<sup>3</sup> V základním prostředí R musíme po napsání příkazu stisknout Enter



**Obr. 3. Obrazovka po spuštění R commanderu**

Zatímco základní prostředí R má jen jedno aktivní okno, kam se píše příkazy a poté se zde generují i výstupy, případně se vypisují i chybová hlášení, v R commanderu je vše poměrně přehledně rozděleno. Nahoře je poměrně bohatá sada nabídek (R commander předpokládá, že vše se bude provádět skrze nabídku, je ale možné užívat i příkazy). Příkazy je možné psát (resp. R commander je tam vypíše po volbě příslušných nabídek) do horního velkého okna. Spodní velké okno pak slouží ke generování výstupů (typicky tabulek). Spodní okraj je pak určen k vypisování informačních a chybových hlášek (barevně odlišeny, chybové hlášky jsou červeně, informační modře a zeleně). Dodejme, že grafy (pokud je vygenerujeme) se ale spustí do základního prostředí R, tedy v R commanderu je neuvídíme a musíme se přepnout do základního okna R.



## Představení analytických možností R commanderu

Představíme si nejdříve základní nabídky v R commanderu, poté si některé ukážeme detailněji. Základní struktura nabídek je následující:<sup>4</sup>

**Data** – otevření a import dat, úpravy dat, uložení dat (R formát či textový formát)

**Statistics** – základní statistické procedury, od popisné statistiky po regresní (vč. logistické) a faktorovou analýzu (vč. konfirmační)

**Graphs** – základní statistické grafy – histogram, krabíčkový, bodový

**Distributions** – možnost hledání kvantilů a generování výběrů z běžně známých rozdělení (N, t, F aj.)

**Tools** – možnosti nastavení a načítání balíčků

### Nabídka Data a její možnosti

Skrze nabídku Data si můžeme jednak vytvořit vlastní datový soubor, jednak je možné otevřít stávající data z různých formátů. Pokud chceme vytvořit nový datový soubor zvolíme z nabídky možnost *Data-New data set*. Poté musíme určit jméno datového souboru a poté můžeme vkládat data do tabulky. Na konci naší práce vybereme *Exit and save*. Musíme být ale obezřetní, soubor se uloží jen do paměti, pro jeho trvalé uložení pak musíme ještě vybrat nabídku *Data-Active data set-Save*. Zde volíme formát pro uložení dat, v R je nativní formát tzv. Rdata, lze ale vybrat i txt soubor (skrže *Data-Active data Export*). Již nyní upozorníme na skutečnost, že v R (i R commanderu) můžete pracovat s více soubory najednou. V R commanderu se aktivní soubor vybírá nahoře vlevo na liště. Obecně práci s více soubory najednou spíše nedoporučujeme.

Spíše než pro tvorbu vlastních dat využijeme R (a R commander) pro zpracování existujících dat. Samozřejmě lze načíst nativní formát (Rdata), ale skrze různé balíčky (už předinstalované v base modulu) lze načítat i jiné datové formáty: XLS (Excel), SAV (SPSS), B7DAT

---

<sup>4</sup> Vynecháváme File, Edit a Help.

(SAS), dále i Stata a Minitab. Pro načtení dat využijeme nabídku *Data-Import data* a průvodce, který je zde obsažen. Po načtení dat je vždy vhodná kontrola. Jednak na stavové liště (zcela dole) si můžeme přečíst počet případů v načtených datech), jednak je vhodné zkontrolovat aspoň některé z načtených proměnných (např. skrze četnostní tabulku, srov. dále).

Další části nabídky *Data* umožňují provádět některé datové operace prováděné před samotnou analýzou, případně umožňují přípravu nových proměnných, Volba *Data-Active data set* nabízí prohlížení dat, třídění dat, agregaci dat, výběr části dat, uložení a export dat (viz výše). Dílčí nabídka *Data-Manage variables in active data set* pak slouží pro datové transformace, konkrétně pro výpočet nové proměnné, rekodování, standardizaci, či dichotomizaci.

### Nabídka Statistics a její možnosti

R rozlišuje mnoho typů proměnných. R commander v zásadě rozlišuje dva základní typy, buď je proměnná číselná a bude s ní nakládáno jako z kardinální, nebo je kategoriální (slovy R commanderu factor) a bude s ní nakládáno jako s ordinální proměnnou. To znamená, že nám R commander neumožní například výpočet průměru, mediánu. Naopak pro číselnou proměnnou nelze například generovat kontingenční tabulku. Každá procedura nám tak nabízí jen ty proměnné, které svým typem této proceduře odpovídají.

#### I. Nabídka *Statistics-Summaries*

**Active data set** – podává základní přehled proměnných v souboru, pro číselné určí průměr, maximum a minimum, pro kategoriální (typ factor) zobrazí četnostní tabulku

**Numerical Summaries** – základní popisné statistiky číselných proměnných (viz záložka Statistics)

**Frequency distributions** – četnostní tabulka kategoriálních proměnných (plus chi-kvadrát test rozložení četností – default rovnoměrné rozložení)

**Table of Statistics tabulka s průměry (či mediány)** pro číselné proměnné v třídění dle jedné kategoriální proměnné



## II. Nabídka *Statistics-Means*

**T-testy:** jedno, dvouvýběrový a párový – pro číselné proměnné, slouží k porovnání průměrů.

**Analýza rozptylu** (včetně následných testů včetně grafů CI)

### **Vícefaktorová analýza rozptylu**

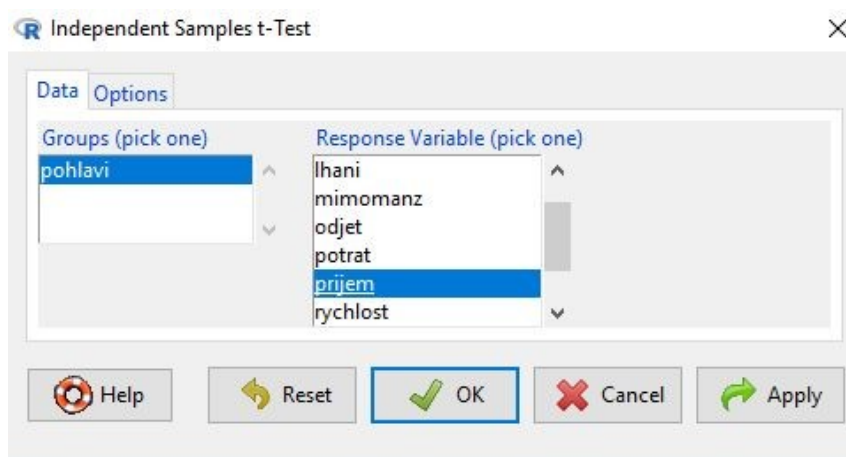
Upozornění: V rámci t-testu ani analýzy rozptylu není obsažen test shody rozptylů (viz *Statistics-Variances*)

### **Praktický příklad č.1** (dvouvýběrový t-test)

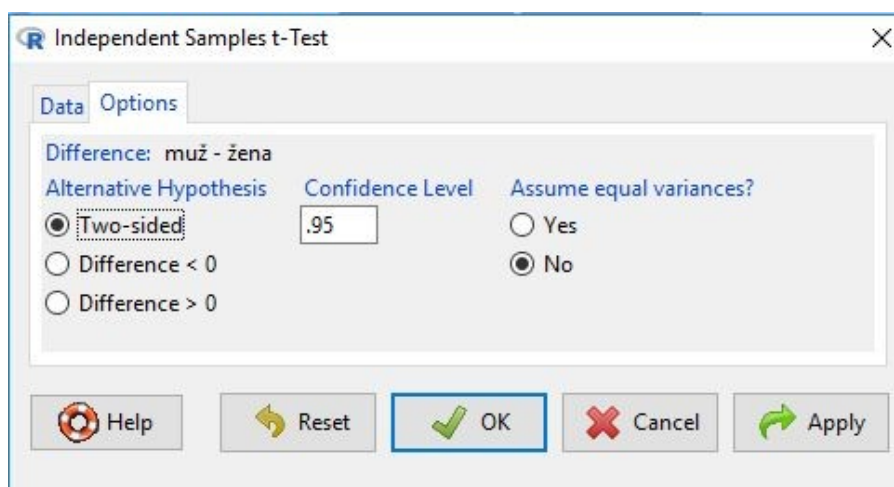
Nabídka *Statistics-Means-Independent sample t-test*, data EVS99-cast.sav

1. Po výběru dialogu na dvouvýběrový t-test vybereme proměnné, budeme srovnávat průměrný příjem mužů a žen (proměnné *prijem* a *pohlavi*)
2. Dialog zobrazuje obrázek 4 (proměnná *prijem* je zadána jako Response variable, proměnná *pohlavi* jako Groups).
3. V rámci volby *Options* (obr. 5) lze volit oboustranný či jednostranný test a určit, zda máme ve skupinách stejné rozptyly. Ponechám oboustranný test a obecnější variantu pro neshodné rozptyly (tzv. Welchův test). Pokud bychom chtěli detekovat (ne) shodu rozptylů, museli bychom ještě před t-testem provést test shody rozptylů (např. Leveneho testem v

**Obr. 4. Základní dialog pro zadání dvouvýběrového t-testu (výběr proměnných)**



Obr. 5. Dialog pro volby v rámci dvouvýběrového t-testu



R commander nabízí jen základní výstupy – průměry, směrodatné odchylky, interval spolehlivosti a statistický test (testové kritérium, stupně volnosti a P hodnotu):

```
Welch Two Sample t-test
```

```
data: příjem by pohlaví
```

```
t = 2.6262, df = 1098.8, p-value = 0.008754
```

```
alternative hypothesis: true difference in means is not equal  
to 0
```

```
95 percent confidence interval:
```

```
272.6577 1883.8644
```

```
sample estimates:
```

```
mean in group muž mean in group žena
```

```
14535.17
```

```
13456.91
```

**Stručná interpretace výsledků:** Průměrný příjem mužů v našem souboru (14535) je cca o 1 tisíc větší než u žen (13457). Nulová hypotéza testu říká, že v populaci (dospělí ČR) bude průměrný příjem mužů a žen stejný, alternativa (oboustranná) pak tvrdí, že mezi těmito skupinami existuje rozdíl. Na základě průměrů obou skupin a jejich směrodatných odchylek se

vypočte testové kritérium ( $t = 2.6262$ ), které má v našem případě počet stupňů volnosti  $df=1098,8$ . Výsledná P hodnota je poměrně malá ( $p\text{-value} = 0.008754$ ). Při klasickém postupu (P je menší než dopředu stanovená mez, typicky 5%) pak tedy zamítneme nulovou hypotézu a přijmeme hypotézu alternativní, tj. konstatujeme, že mezi muži a ženami je v průměrných příjmech v ČR rozdíl. Interval spolehlivosti (95 percent confidence interval: 272.6577 1883.8644) nám pak dává hrubou představu o tom, jak velký rozdíl mezi průměrnými příjmy mužů a žen může v populaci být.

**Upozornění:** R commander nepočítá např. míry věcné významnosti ( $d$ ,  $\text{Eta}^2$ ) je třeba dopočítat ručně, nebo užívat některé jiné balíčky a psát příkazy (více na druhém dni semináře).

### III. Nabídka *Statistics-Proportions a Nonparametric Tests*

**Proportions:** Testy pro srovnání proporcí dichotomických proměnných

**Nonparametric Tests** – alternativy k t-testům a ANOVA (M-W test, K-W test) a testy pro párovaná data (2 či více opakování)

### IV. Korelace

Nabídka *Statistics-Summaries-Correlation matrix*

Tato nabídka počítá korelace (Pearson a Spearman) případně vyčíslí P hodnoty (umí i korekci pro vícenásobné testování)

Nabídka *Statistics-Summaries-Correlation test*

Tato nabídka počítá test (i jednostranný), spočte i interval spolehlivosti (opět Pearson, Spearman ale i Kendall)

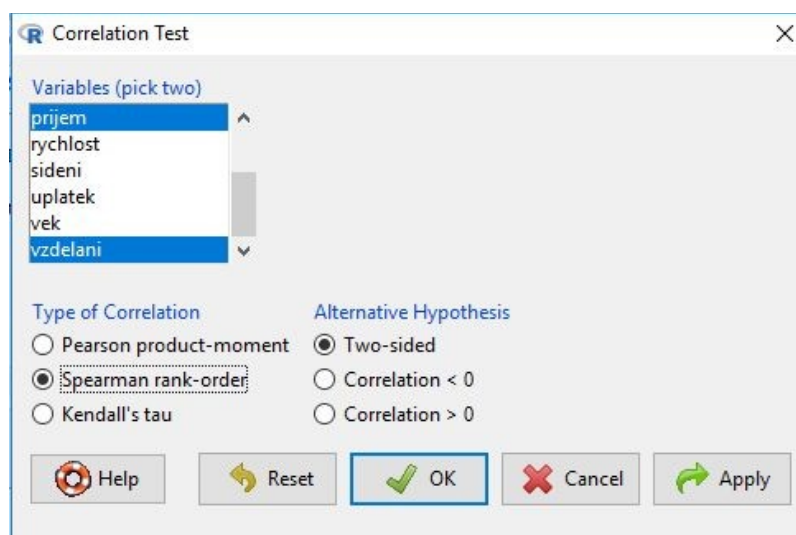


## Praktický příklad č.2 (korelace)

Korelace mezi příjmem (*prijem*) a vzděláním (*vzdelani*).

1. Použijeme proceduru *Statistics-Summaries-Correlation test*.
2. Zadáme dvě proměnné (*prijem* a *vzdelani*). Musíme si uvědomit, že proměnná vzdělání je ordinální (má jen 4 kategorie), volíme tedy některou z neparametrických korelací (zde volíme Spearmanův koeficient-*obr. 6*).
3. Ponecháme volbu oboustranného testu (je ale možné zvolit i jednostranný, který očekává, že půjde o pozitivní korelaci – tj. s růstem vzdělání roste příjem).

Obr. 6 Základní dialog pro zadání korelace vč. testu



R commander nabízí klasické výstupy: tj. hodnotu korelace, testové kritérium a P hodnotu:

```
Spearman's rank correlation rho
```

```
data:  příjem and vzdělání
```

```
S = 159285636, p-value < 2.2e-16
```

```
alternative hypothesis: true rho is not equal to 0
```

```
sample estimates:
```

```
rho
```

```
0.2916606
```



**Stručná interpretace výsledků:** Souvislost příjmu a vzdělání v našem souboru je pozitivní o velikosti 0,29 (rho). Výsledek otestování nezávislosti udává P hodnotu menší než  $2.2e-16$ . Při klasickém postupu (P je menší než dopředu stanovená mez, typicky 5%) pak tedy zamítneme nulovou hypotézu a přijmeme hypotézu alternativní, tj. konstatujeme, že vzdělání a příjem v ČR spolu souvisí.

***Upozornění:***

1. Pokud by naše proměnná ordinálního typu byla definovaná jako factor (tj. například v SPSS datech by měla popisky) R commander by nám neumožnil výpočty korelací.
2. R commander nepočítá intervaly spolehlivosti pro korelace pro Spearmanův a Kendallův koeficient, pro Pearsonův tuto hodnotu vypočte, viz následující výstup (interval spolehlivosti je tučně):

```
Pearson's product-moment correlation
```

```
data:  drogy and euth
```

```
t = 5.8071, df = 1802, p-value = 0.000000007498
```

```
alternative hypothesis: true correlation is not equal to 0
```

```
95 percent confidence interval:
```

```
0.08994718 0.18055746
```

```
sample estimates:
```

```
cor
```

```
0.1355357
```



## V. Regrese

Nabídka *Statistics-Fit models-Linear Regression*

*Základní lineární regresní výstupy:  $R^2$ , F-test a jednotlivé koeficienty vč. T-testů*

Další procedury v rámci nabídky Fit models umí GLM, plus základní modely logistické regrese (pro nominální a ordinální proměnnou)

Skrze nabídku Model lze dále generovat další výstupy: např. AIC, BIC, intervaly spolehlivosti jednotlivých koeficientů)

## VI. *Statistics-Contingency table-Two-way table (nabídka má i tabulku pro 3 proměnné plus proceduru na přímé vložení konti tabulky\*)*

Možnosti jsou omezené jen na vizualizaci tabulky (počty, nebo řádková či sloupcová procenta) a dále na chi-kvadrát test a jeho podrobnosti (příspěvek polí k jeho hodnotě), pro další (konti koeficient, znaménkové schéma) nutno užít pomůcky (např. Excel) či procedury v jiných balíčcích dostupné skrze příkazy

\* Netřeba tedy mít originální data, stačí vložit četnosti a R spočte vše potřebné, tj. chi-kvadrát test)

## VII. *Statistics-Dimensional analysis-Factor analysis (Confirmatory factor analysis)*

Velice snadná zadání EFA či CFA

Omezení u EFA: výstupy jen zcela základní, jen kolmé rotace, nutno ručně volit počet faktorů

CFA poměrně slušné, ale pracuje jen s kardinálními proměnnými (Pearsonova korelace a ML odhad parametrů, umí robustní odhady std. chyb, tj. MLR)





## VIII, Nabídka **Graphs**

Typy grafů se opět nabízí dle typu proměnných (tj. dialog nabízí jen proměnné příslušného typu)

**Histogram** – rozložení četností spojité proměnné (možné žádat i separátní histogramy pro různé skupiny dle jedné kategoriální proměnné)

**Box Plot** – krabičkový graf (možné žádat i separátní krabičky v jednom obrázku pro různé skupiny dle jedné kategoriální proměnné)

**Density estimate** – odhad křivky charakterizující rozložení spojité proměnné (opět lze rozdělit do skupin)

**Plot of means** – graf s průměry spojité proměnné pro různé skupiny dle jedné kategoriální proměnné (včetně intervalů spolehlivosti)

**Scatter Plot** – bodový graf pro dvě spojité proměnné (vhodné pro regresní či korelační analýzu)

Poznámka 1: Grafy se nezobrazují v R commanderu, ale přímo v základním prostředí R, v samostatném okně. Každý nový graf přemaže předchozí, pro další potřebu je tedy vhodné si graf uložit (viz další poznámka).

Poznámka 2: Grafy lze exportovat do PDF nebo jako obrázek, R navíc nabízí různá rozlišení.





### Literatura k R commanderu

Soukup, P., L. Rabušic, P. Mareš.(v přípravě). Statistická analýza sociálněvědních dat v R. Masarykova univerzita.

Fox, J. 2005. The R Commander: A Basic-Statistics Graphical User Interface to R. Journal of Statistical Software. 14(9): 1-42

Fox, J. 2017. Using the R Commander: A Point-and-Click Interface for R. Chapman & Hall/CRC Press

### Literatura k R (česky)

Pekár, S., M. Brabec, 2012 2009. Moderní analýza biologických dat - Zobecněné lineární modely v prostředí R. Scientia.

Pekár, S., M. Brabec, 2012. Moderní analýza biologických dat 2: Lineární modely s korelacemi v prostředí R. MuniPress.

Pekár, S., M. Brabec, 2014. Moderní analýza biologických dat 3: Nelineární modely v prostředí R. MuniPress.

Zvára, K. 2013. Základy statistiky v prostředí R. 1. vydání. Praha : Karolinum.  
[https://www.wikiskripta.eu/w/Kategorie:R\\_wiki](https://www.wikiskripta.eu/w/Kategorie:R_wiki)



EVROPSKÁ UNIE  
Evropské strukturální a investiční fondy  
Operační program Výzkum, vývoj a vzdělávání

