

# Načítání dat do R a ukládání dat na pevný disk

---

17VSADR – Skriptování a analýza dat v jazyce R

Lubomír Štěpánek<sup>1, 2</sup>



<sup>1</sup>Oddělení biomedicínské statistiky  
Ústav biofyziky a informatiky  
1. lékařská fakulta  
Univerzita Karlova v Praze



<sup>2</sup>Katedra biomedicínské informatiky  
Fakulta biomedicínského inženýrství  
České vysoké učení technické v Praze

(2018) Lubomír Štěpánek, CC BY-NC-ND 3.0 (CZ)



Dílo lze dále svobodně šířit, ovšem s uvedením původního autora a s uvedením původní licence. Dílo není možné šířit komerčně ani s ním jakkoliv jinak nakládat pro účely komerčního zisku. Dílo nesmí být jakkoliv upravováno. Autor neručí za správnost informací uvedených kdekoli v předložené práci, přesto vynaložil nezanedbatelné úsilí, aby byla uvedená fakta správná a aktuální, a práci sepsal podle svého nejlepšího vědomí a svých „nejlepších“ znalostí problematiky.

# Obsah

- 1 Pracovní složka
- 2 Import a export dat
- 3 Literatura

# Pracovní složka

- zjištění, která složka je pracovní

```
1 || getwd()
```

- výpis obsahu pracovní složky formou vektoru

```
1 || dir()
```

- nastavení pracovní složky

```
1 || setwd("C:/.../my_working_directory")
```

- nastavení pracovní složky dialogovým oknem

```
1 || setwd(choose.dir())
```

# Import a export volného textu

- pomocí funkcí `readLines()` a `writeLines()`
- lze tak nahrát libovolná data, která mají textovou reprezentaci

```
1 my_html <- readLines(  
2   con = paste(  
3     "https://predmety.fbmi.cvut.cz",  
4     "cs/17VSADR",  
5     sep = "/"  
6   ),  
7   encoding = "UTF-8"  
8 )
```

# Import a export volného textu

- uložení a načtení desktopového dokumentu

```
1      writeLines(                # ukládám textový dokument
2          text = paste(
3              "One R to rule them all",
4              "one R to find them",
5              "one R to bring them all",
6              "and in the darkness bind them",
7              sep = "\n"          # separátor typu nový řádek
8          ),
9          con = "my_text.txt"
10     )
11
12     my_loaded_text <- readLines(
13         con = "my_text.txt",
14         encoding = "UTF-8"
15     )                                # načítám textový dokument
```

# Import a export dat tabulky

- zánění funkcí je `read.table()` a `write.table()`
- obě funkce mají řadu wrapperů ( `read.csv()` a `write.csv()`, `read.delim()` a `write.delim()` a další)

```
1      write.table(                                # ukládám data.frame
2          x = mtcars,
3          sep = ";",
4          row.names = FALSE,
5          file = "mtcars.csv" # anebo "mtcars.txt"
6      )
7
8      my_mtcars <- read.table(
9          file = "mtcars.csv",
10         sep = ";",
11         header = TRUE
12     )                                           # načítám data.frame
```

# Import a export dat tabulky

- funkce `read.table()` má spoustu užitečných argumentů

```
1 write.table(iris, "iris.txt")
2 my_iris <- read.table(
3     file = "iris.txt",
4     sep = " ",
5     header = TRUE,
6     stringsAsFactors = FALSE,
7     nrows = 1,      # načte jen první řádek;
8                     # může se hodit pro odhad koerce,
9     check.names = FALSE,
10                    # vynechá kontrolu korektnosti
11                    # popisků sloupců
12     colClasses = "character"
13                    # přetypuje všechny sloupce
14                    # na textové proměnné
15 )
```



# Import a export dat tabulky z MS Excel® (.xlsx)

- vhodný je například balíček openxlsx
  - má výhodu, že narozdíl od balíčku např. xlsx nepotřebuje Java Tool Kit, takže dokáže najednou nahrát více souborů MS Excel®
- uložení tabulky do excelového formátu (.xlsx)

```
1 | browseURL(paste(  
2 |   "https://raw.githubusercontent.com/LStepanek",  
3 |   "17VSADR_Skriptovani_a_analyza_dat_v_jazyce_R",  
4 |   "master/export_dat_do_ms_excelu.R", sep = "/"  
5 | ))
```

- načtení excelové tabulky

```
1 | my_data<- read.xlsx(  
2 |   xlsxFile = "moje_tabulka_je_ted_v_excelu.xlsx",  
3 |   sheet = 1,      # anebo jméno listu  
4 |   colNames = TRUE  
5 | )
```

# Import bitmapového obrázku

- vhodný je například balíček png, jpeg či raster

```
1 my_picture <- readJPEG(  
2   "__03_landmarky__.jpg"  
3 )
```

- výsledkem je array o třech rozměrech
  - svislá souřadnice
  - vodorovná souřadnice
  - barevné kanály

# Export konzolového výpisu do souboru

- pomocí `sink()` - `sink()` anebo `capture.output()`

```
1 | (muzi <- rnorm(100, mean = 175, sd = 10))
2 | (zeny <- rnorm(100, mean = 160, sd = 10))
3 |
4 | t.test(muzi, zeny)
5 |
6 | # výpis z konzole do textového souboru
7 | capture.output(
8 |     t.test(muzi, zeny),
9 |     file = "t_test.txt"
10 | )
11 |
12 | # anebo
13 | sink("tohle_je_taky_t_test.txt")
14 | t.test(muzi, zeny)
15 | sink()
```

# Export smysluplného R-kového objektu do T<sub>E</sub>X-ového kódu

- pomocí balíčku xtable

```
1      library("xtable")
2
3      my_linear_model <- lm(mpg ~ hp + cyl,
4                             mtcars)
5
6      print(
7          xtable(my_linear_model,
8                  digits = 4),
9          floating = FALSE,
10         tabular.environment = "tabular",
11         hline.after = NULL,
12         include.rownames = TRUE,
13         include.colnames = TRUE
14     )
```

# Import „exotických“ souborů do R

- naším přítelem je balíček `foreign`
- podporuje načítání dat z formátů
  - Epi Info
  - Minitab
  - S
  - SAS, SPSS, STAT, Systat, Weka

```
1 library("foreign")
2
3 # import dat z SPSS
4 my_data <- read.spss(
5     file = "du1_30.sav",
6     to.data.frame = TRUE
7 )
```

# Intermezzo

- jak importovat data z MS Word®?
- jak načíst do R datovou strukturu list?

# Stažení HTML obsahu webové stránky do R (webscraping)

- načítání volného textu s kódováním

```
1 | my_html <- readLines(  
2 |   con = "https://predmety.fbmi.cvut.cz/cs/17  
   vsadr",  
3 |   encoding = "UTF-8"  
4 |   # v úvahu připadá "latin1", "latin2", "ASCII"  
5 | )
```

- náhrada pevných mezer vlastním tagem

```
1 | my_text <- gsub("&nbsp;", "M_E_Z_E_R_A",  
2 |               my_html)
```

- vymazání všech HTML tagů a entit

```
1 | my_text <- gsub("<.*?>", "", my_text)  
2 | my_text <- gsub("&.*?;", "", my_text)  
3 | my_text <- gsub("M_E_Z_E_R_A", " ", my_text)  
4 |   # nahrazuji nezlomitelnou mezeru zlomitelnou
```

# Stažení HTML obsahu webové stránky do R (webscraping)

- odstranění white spaces

```
1 | for(i in 1:length(my_text)){  
2 |   while(grepl(" ", my_text[i])){  
3 |     my_text[i] <- gsub(" ", " ", my_text[i])  
4 |   }  
5 | }
```

- ponechání jen řádků obsahujících text

```
1 | my_text <- my_text[!my_text %in% c("", " ")]
```

- jednoduchá úloha – extrakce emailů

```
1 | gsub(  
2 |   "(.*?)([a-zA-Z\\.]+@[a-zA-Z]+\\.([a-zA-Z]+)(.*)"  
3 |   ,  
4 |   "\\2",  
5 |   my_text[grepl("@", my_text)]  
6 | )
```



# Odstranění diakritiky ve staženém textu

- někdy se může hodit odstranění diakritiky

```
1 my_text <- iconv(  
2  
3   my_text,  
4   from = "UTF-8",           # nepovinný argument  
5   to = "ASCII//TRANSLIT"  
6  
7 )
```

# Literatura



Karel Zvára. *Základy statistiky v prostředí R*. Praha, Česká republika: Karolinum, 2013. ISBN: 978-80-246-2245-3.



Hadley Wickham. *Advanced R*. Boca Raton, FL: CRC Press, 2015. ISBN: 978-1466586963.

Děkuji za pozornost!

lubomir.stepanek@lf1.cuni.cz

lubomir.stepanek@fbmi.cvut.cz

► GitHub

`github.com/LStepanek/17VSADR_Skriptovani_a_analyza_dat_v_jazyce_R`