R - úvod do problematiky

Lubor Homolka

8. května 2014

Obsah školení

Teoretická část:

- Reproducible research
- R $\sim$ Open-science filosofie

Praktická část:

- Literate Coding in R
- Zdroje a správa dat
- Vizualizace dat jako příprava pro inferenční statistiku

Reproducible research

- REsearch
- Proč je Reproducible research aktuální téma?
- váha důkazu $\sim$ p-value?

Literate Coding I

Záznam z experimentu

Cílem *experimentu* bylo ...
a postupovali jsme následovně:

„Naměřili jsme hodnoty 3,4 a 5 a z nich jsme vypočítali průměr (prostý aritmetický)“

Očekávali jsme, že průměr bude 2, ale nám vyšel 4, což se ale při počtu pozorování dalo očekávat.

Literate Coding II

Záznam z experimentu

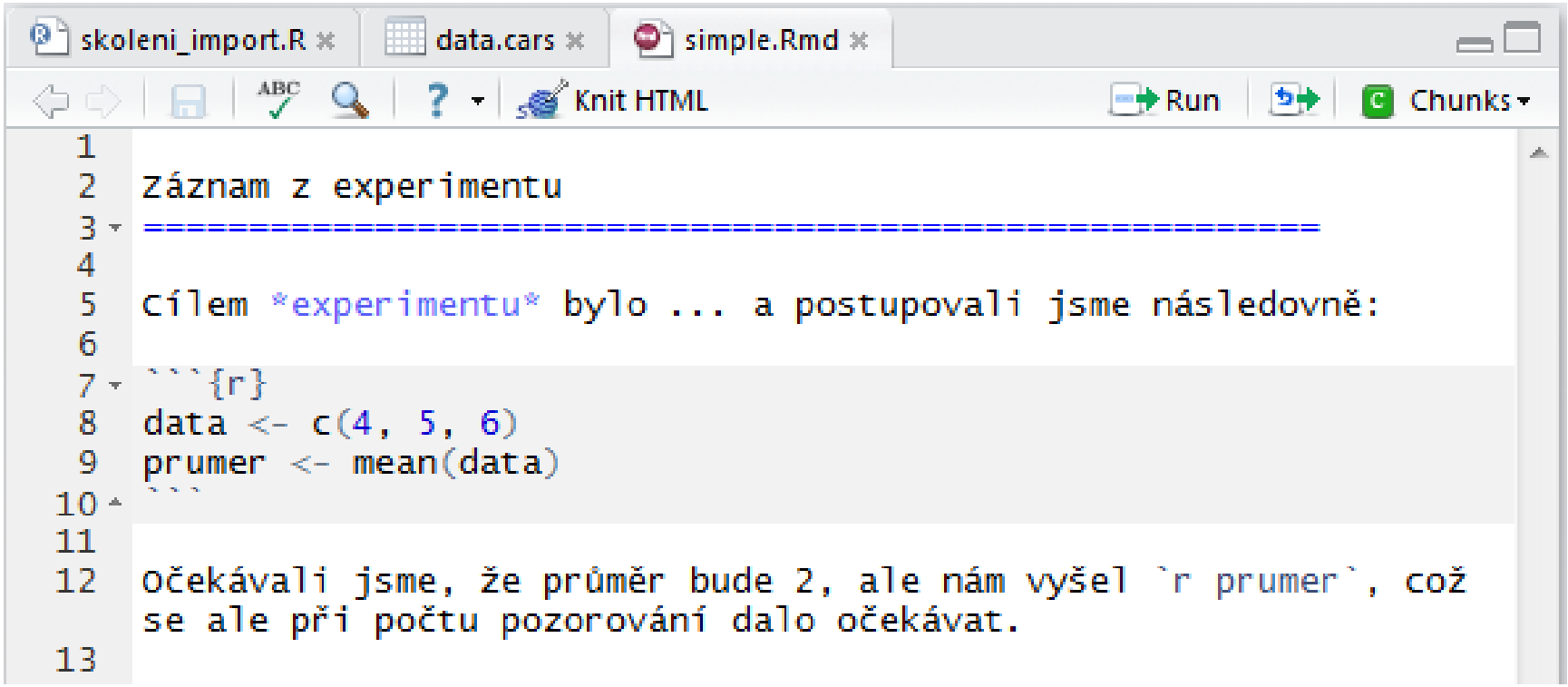
Cílem *experimentu* bylo ...
a postupovali jsme následovně:

„Naměřili jsme hodnoty 3,4 a 5 a z nich jsme vypočítali průměr (prostý aritmetický)“

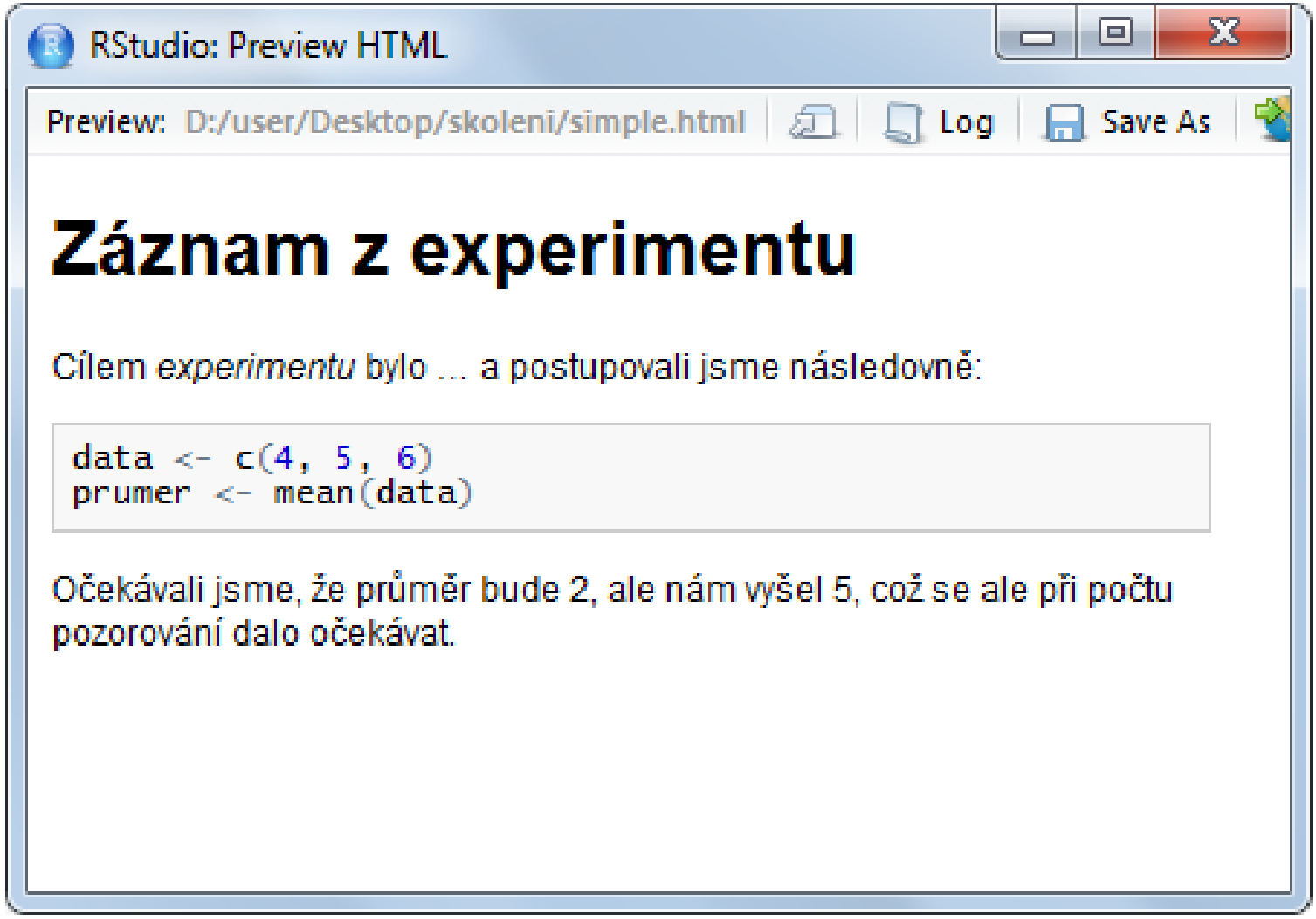
Očekávali jsme, že průměr bude 2, ale nám vyšel 4, což se ale při počtu pozorování dalo očekávat.

Běžný text, který je možné psát jazyky: HTML, Markdown nebo LaTeX.
„Počítačový jazyk“ - v našem případě R

Literate Coding III - Knitr



Literate Coding IV - Výsledek



Lokální data I

- Ideální zdroj je textový soubor bez *zbytečného* formátování.
- preferované typy: `.txt` nebo `.csv`

Přihlašte se na
`http://www.fame.utb.cz/pokr/`
studijni_materialy -> Podniky -> r_skoleni

a stáhněte si na disk (pracovního adresáře) soubor `cars.txt`

Načtěte data do R příkazem:

```
data.cars <- read.table("cars.txt",sep=" ", header=TRUE)
str(data.cars) #struktura dat
```

Lokální data II

Další zdroje: `.xlsx` nebo `.acddb`

```
library(xlsx)
excel.data <- read.xlsx("analyza1.xlsx", sheetIndex=1)
```

Data z Internetu I

R umožňuje práci se vzdálenými zdroji, jak „pod heslem“, někdy ale problematické https.

```
url.cars<-"http://www.stat.ucla.edu/~jeroen/ggplot2/mtcars.txt"
data.cars <- read.table(url.cars,sep=" ", header=TRUE)
write.table(x=data.cars, file="auta.txt", row.names=FALSE)
```

Data z Internetu II – quantmod

```
casove.rady <- new.env()
start.date = as.Date("2010-01-11")
end.date = as.Date("2014-05-04")
akcie <- c("GOOG","UKX")

getSymbols(akcie, env = casove.rady,
 src = "yahoo",
 from = startDate,
 to = endDate)

head(casove.rady$GOOG)
tail(casove.rady$UKX)
plot(casove.rady$UKX)
barChart(casove.rady$UKX)
```

Manipulace s daty

S daty je možné dělat snad úplně vše i v `base` knihovně. Práce ale není ani příliš rychlá, ani „elegantní“.

Základní funkce:

- `*apply` a `order`
- `d*ply`, zejména `ddply`

Vizualizace dat

`R` využívá několik knihoven k práci s grafikou. Představíme si základní (base) a dvě rozšíření:

- `lattice`
- `ggplot2`

Prezentace a její obsah je samozřejmě reproducible!

`https://github.com/luboRprojects`