Úvod do R

Peter Kvasnička

Univerzita Karlova, Praha

Kurz pre 4. ročník BMF
Jeseň 2017

Čo sa naučíte v tomto kurze

1. Používať R

Trocha neskromné, nemáme veľa času.

Načítať dáta import z textových súborov a Excelu

Preskúmať dáta kreslenie - `ggplot2`

Upraviť dáta manipulácia s dátami, súhrny, filtrovanie atď. - `tidyverse`

Analyzovať dáta lineárny model, ANOVA atď.

Validovať analýzu simulácia, replikácia (bootstrap)

Nie presne v tomto poradí, budeme sa hýbať v kruhoch.

Čo sa naučíte v tomto kurze

2. Používať moderný ekosystém pre prácu s dátami

Open source software, nepotrebujeme nakupovať drahý štatistický software, ani MS Office, všetko čo potrebujeme sa dá stiahnuť z Internetu.

Version control Chceme, aby sa dáta dali zdieľať a boli chránené pred stratou alebo nechcenou zmenou.

Rôzne zdroje dát Chceme pracovať s dátami z viacerých možných zdrojov - textové súbory, Excel, JSON, databázy atď.

Zdieľanie dát a analýzy Chceme, aby si ľudia mohli skontrolovať našu analýzu.

Outline

Čo sa očakáva, že budete vedieť

Základné znalosti z pravdepodobnosti a štatistiky

- Stačí, aby ste sa veľmi nezľakli, keď poviem t-test.
- Máte určitú prax v spracovaní a zobrazovaní dát.

Programovanie

- Očakávam, že máte za sebou kurz programovania, v hocičom.
- Napríklad keď idete písať kód, začnete tým, že si zapnete anglickú klávesnicu.
- A pochytili ste trocha algoritmického myslenia.
- Budeme sa učiť nový jazyk a používať nové nástroje, takže pôjdeme od nuly.

Angličtina

- Je mi ľúto, ale bez angličtiny budete mať v tomto kurze ťažkosti.
- Predovšetkým si veľmi ťažko budete hľadať pomoc na Internete, a to je prvá vec, ktorú človek robí, keď mu niečo nefunguje alebo nevie, ako niečo urobiť.

Ešte potrebujete laptop

Laptop

- Windows 7 alebo 10, alebo Linux
  - Windows 10 má WSL - Windows Subsystem for Linux - a umožňuje vám lepšie používať niektoré veci, napríklad `git`.
  - Ale Linux je na to ešte lepší.
- Nepotrebujete mať extra silný procesor alebo veľa pamäti, aspoň nie pre tento kurz.

Inštalovaný software

- Potrebujete mať nainštalované R a RStudio.
- Zriaďte si účet na GitHub (`https://www.github.org`) a stiahnite si *GitHub Desktop*. Ak máte Linux, stačí vám nainštalovať git.
  - Toto stačí v priebehu kurzu, svoj GitHub account budete potrebovať na odovzdanie zadaní.
  - Ale nezaškodí získať trocha praxe, takže čím skôr, tým lepšie.
- Návod na inštaláciu R/RStudio a na zriadenie GitHub konta nájdete v GitHub repozitári tohoto kurzu, `https://www.github.com/PKvasnick/RTutorial/`
- Ak by ste mali problémy s inštaláciou, rýchlo sa ozvite.

Kde hľadať informácie

Príručky

Príručiek je veľa, väčšina aktuálnych a moderných je v angličtime.

- Na Pinterestovej stránke `https://sk.pinterest.com/peterkvasnika/my_r/` nájdete odkazy na niekoľko internetových portálov a PDF dokumentov, ktoré vám môžu pomôcť v začiatkoch.
  - Niekoľko z nich je slovenčine/češtine.
- Na portáli CRAN (Comprehensive R-Archive Network - `https://cran.r-project.org/`) nájdete prehľad dokumentácie k R.

Nápoveď v R a RStudio

- R má svoj vlastný help systém, naučíte sa s ním pracovať.
- RStudio má takisto svoje helpy.

Kde hľadať informácie

Internet

- To čo programátor robí najčastejšie je, že vysvetlí Googlu lámanou angličtinou čo chce urobiť (`R create dataframe`), alebo priamo do riadku vyhľadávača skopíruje chybovú hlášku.
- S vysokou pravdepodobnosťou nájdete použiteľnú odpoveď, či už je vaša otázka triviálna alebo zložitá.
  - Tú odpoveď nájdete najčastejšie na webe StackOverflow, `https://stackoverflow.com`, s ktorým sa určite spriatelíte.
- Časom prídete na to, že kúsok fungujúceho kódu býva užitočnejší ako podrobný výklad syntaxe.

Prečo R?

Máme predsa ...

Excel a iné tabuľkové programy

SPSS Statisticu, Minitab a iné komerčné programy poskytujúce analýzu na kľúč

Tak prečo mám používať niečo, čo sa treba určitý čas učiť?

Niekoľko dôvodov

Excel nie je štatistický program

- Excel je výborný nástroj na vkladanie dát, získavanie a konsolidáciu dát z databáz a na základné úpravy dát
- Ale nie je dobrý na výmenu dát (polo-proprietárny formát - nikdy nevieme, kedy sa zmení)
- Vzorce v bunkách sa ťažko spravujú a neexistuje praktický spôsob, ako nezávisle dokumentovať, čo sa ako počíta.
- Nemáme výstrahu, ak náhodne zmeníme obsah bunky
- Nástroje pre štatistiku sú implementované ledabylo.
- Grafy sú na zaplakanie.

Niekoľko dôvodov

Robíme stále zložitejšie analýzy

- Chceme skúmať analyzovať zložité a veľké dáta
- Chceme validovať našu analýzu pomocou simulácií a replikácie - potrebujeme analýzy opakovať tisíckrát
- Chceme formulovať a testovať zložité modely (*Data Science*)

Chceme zdieľať dáta a analýzu

- Potrebujeme otvorený software a nie drahé štatistické balíky alebo MS Office
- Potrebujeme otvorené formáty dát
- Chceme software, ktorý sa *rýchlo inovuje*
- Chceme software, ktorý je správny

Preto chceme R!

R je programovate2né

- R je interpretovaný programovací jazyk
- R podporuje integráciu s inými programovacími jazykmi - môžeme volať funkcie naprogramované v C++, Fortrane ap., čo podstatne kompenzuje pomalosť vlastného interpreta R.
- R spolupracuje s Pythonom, Javou a ďalšími jazykmi, ktoré používajú vývojári v data science
- R má výborné IDE, RStudio, a najnovšie aj Visual Studio!

R má bohaté rozhrania pre dáta

- R umožňuje čítať dáta z veľkého množstva vstupných formátov:
  - textových súborov
  - Excelu
  - JSON
  - databáz
  - Apache Spark-u
  - ...
- R dokáže dáta, grafy a reporty exportovať do veľkého množstva formátov.

Preto chceme R!

R je rozšíriteľné

- To, čo robí R skutočne cenným, je ekosystém rozšírení - baličkov (packages)
- Tieto baličky obsahujú všetky štatistické metódy, ktoré kedy budete potrebovať
- Balíčky sídlia na serveri CRAN, a môžete si ich ľahko doinštalovať cez interpret R (`install.packages(<menobalíčka>`
- Balíčky neustále pribúdajú: Ak niekto opublikuje novú štatistickú metódu, s veľkou pravdepodobnosťou ju hneď implementuje v R.

R je renomované a spoľahlivé

- Pretože R používa veľa ľudí, je dobre otestované a všetky prípadné chyby sú hneď odstránené.
- Ak si svoje dáta analyzujete v R, nikto sa nebude pýtať, či ste správne počítali ANOVu.
- Kód vašej analýzy je univerzálne zrozumiteľný doklad o tom, čo ste robili.

Nemá chybu...?

R má svoje špecifiká a slabé stránky

- R sa pôvodne vyvinulo z funkcionálneho a objektovo-orientovaného jazyka S. Preto niektoré veci pracujú trocha odlišne.
- Pretože R má za sebou dlhú históriu, obsahuje niekoľko súperiacich koncepcií a funkčných rozhraní. Preto niektoré veci možno robiť rôznymi spôsobmi, a naopak niektoré podobné veci musíte robiť odlišne.
- R je pomalé. Treba sa vyhýbať zložitým programovým konštrukciám v R (cyklom `for` a podobne), a používať čo najviac metafunkcie R (`apply`), aby sa počítanie robilo v C a Fortrane, a nie v R.
- Napriek tomuto všetkému sa základy programovania v R možno naučiť pomerne rýchlo a pomerne rýchlo získať výsledky.

Ideme na to...

- Otvorte si v prehliadači stránku `https://www.github.com/PKvasnick/RTutorial/` a stiahnite si z adresára `code` všetky súbory `*.Rmd`.
  - Preklikajte sa k súboru a zvoľte *Raw* zobrazenie.
  - Pravý klik a *Save As...*. Je to čisto textový súbor.
- Odporúčam vytvoriť si podobnú adresárovú štruktúru ako v mojom repozitári.
- (Úplne najlepšie) Môžete si tiež vytvoriť klon môjho repozitára pomocou GitHub Desktopu.
- Spustite RStudio, `File->Open...` a nájdite súbor `.R01_PrveKroky.Rmd`