Analiza datelor în R

Curs 1

Despre R

- program open source pentru calcul statistic
- bazat pe limbajul S (J. Chambers, 1976)
- dezvoltat de R. Gentleman şi R. Ihaka (1993); în prezent - mii de contribuitori
- interfaţă bazată pe linii de comandă $\rightarrow$ flexibilitate
- tehnici şi facilităţi pentru:
  - organizare de date
  - grafică
  - calcul numeric
  - inferenţă statistică şi modelare
  - simulare

Resurse utile

1. G. James, D. Witten, T. Hastie, R. Tibshirani - *An Introduction to Statistical Learning with Applications in R*, Springer, 2013.
2. T. Fischetti - *Data Analysis with R*, Packt Publishing, 2015.
3. W. J. Braun, D. J. Murdoch - *A first course in statistical programming with R*, Cambridge University Press, 2007.
4. J. M. Chambers - *Software for Data Analysis. Programming with R*, Springer, 2008.
5. Y. Zhao - *R and Data Mining: Examples and Case Studies*, `http://www.rdatamining.com`

Consola R

Instalare

`https://www.r-project.org`

Directorul curent

- identificare: `getwd( )`
- modificare: `setwd(...)` sau *File* → *Change dir...*
- listarea obiectelor din workspace: `objects()`

Instrucţiuni

- se introduc după Command prompt (>) şi se execută cu ENTER
- o comandă nefinalizată se poate continua pe linia următoare, care va începe cu "+"
- mai multe comenzi pe o linie se separă prin ;
- comenzile pot fi salvate într-un script şi rulate apoi cu Ctrl+R.
- comentariile se pun după #.

Consola R

- Operaţii aritmetice: +, - , * , / ,$\widehat{}$, %%, %/%
- Asignarea se face cu <- sau =.
- Numele variabilelor pot conţine litere, cifre şi . şi încep întotdeauna cu o literă. (Case sensitive!)
- De evitat numele rezervate (q, c, T, F etc).
- Implicit, R afişează 7 cifre semnificative. Acest lucru se poate modifica utilizând

```
options(digits=x)
```

- Operatori logici: &, |, !
- Operatori relaţionali: $==, <, >, <=, >=, !=$

Consola R

- Funcţii matematice predefinite: sin(), cos(), tan(), exp(), log(), sqrt(), floor(), ceiling(), etc
- Constante predefinite e.g. pi.
- help:

```
?nume_functie
```

```
args(nume_functie)
```

- redirecţionarea output-ului către un fişier:

```
sink("numefisier.txt")
...
sink()
```

Vectori

- se definesc cu ajutorul funcţiei `c(...)` ("concatenate")
- elementele se separă prin virgulă
- componentele sunt de acelaşi tip, numeric, logic (`T=TRUE, F=FALSE`) sau caracter
- selectare elemente:

```
a[i]
a[i:j]
a[c(i1, i2, ..., ik)]
a[-i]
a[-c(i1, i2, ..., ik)]
```

- numărul de componente se determină cu

```
length(nume_vector)
```

Vectori

Modalităţi de definire automată:

- `a:b` $\rightarrow a, a+1, a+2, \ldots, \leq b$
- `seq(a, b, by=pas)`
- `seq(a, b, length=k)`
- `numeric(n)` $\rightarrow$ vector de lungime n cu toate componentele 0
- `rep(n,k)` $\rightarrow$ vector în care n se repetă de k ori

Operaţii cu vectori

- Operaţiile cu vectori (+, -, *, /, $\hat{}$) se efectuează element cu element.
- Variabilele numerice sunt vectori de lungime 1.
- : este executat înainte de +, -, *, /, dar nu şi de $\hat{}$.
- Atunci când o operaţie implică doi vectori de lungimi diferite, cel mai scurt este "reciclat" de oricâte ori este necesar.
- Compararea a doi vectori se face element cu element. Rezultatul este un vector de valori logice corespunzătoare comparaţiilor individuale.

Exerciţii

1. Să se genereze următoarele şiruri:
   i) 1, 3, 5, . . . , 999.
   ii) 1, 1, 1, 1, 2, 2, 2, 2, 3, 3, 3, 3
   iii) 1, 1, 1, 2, 2, 3, 3
   iv) 1, 2, 3, 4, 5, 6, 7, 6, 5, 4, 3, 2, 1
   v) 1, 1/2, 1/3, . . . , 1/10
   vi) 1, 8, 27, 64, 125, 216

Exerciţii

2. Să se calculeze şi să se discute:

i) `seq(0,10.5,by=1)`

ii) `-0.5:10`

iii) `0:10-0.5`

iv) `seq(0.5, 9.5)`

v) `10:22/10`

vi) `10/2:22`

vii) `(10/2):22`

viii) `(1:2)*(0:3)`

ix) `r=1:5; s=-2:2; s/r; r/s; s/s`