ТЕОРИЈА УЗОРАКА час 1

5. март ’14.

УВОД У R

- “R *is really important to the point that it’s hard to overvalue it. It allows statisticians to do very intricate and complicated analyses without knowing the blood and guts of computing systems.*”

  D. Pregibon, научни истраживач у Google-у

  The New York Times , 2009.

- *Names You Need to Know in 2011:*

  R *Data Analysis Software*

  Forbes

- Предности:
  - слободан (*open source*) компјутерски програм, издат под GPL лиценцом
  - комплетан програмски језик, у оквиру кога су доступне функције за манипулацију подацима и извршење великог броја статистичких и нумеричких метода
  - окружење, које омогућава високо квалитетну визуелизацију података и коришћење многобројних графичких алата
  - флексибилност и прилагодљивост конкретном проблему
  - снабдевеност богатом колекцијом пакета, која се непрекидно развија и проширује
  - могућност интеграције R кода унутар LaTeX-а
  - ...

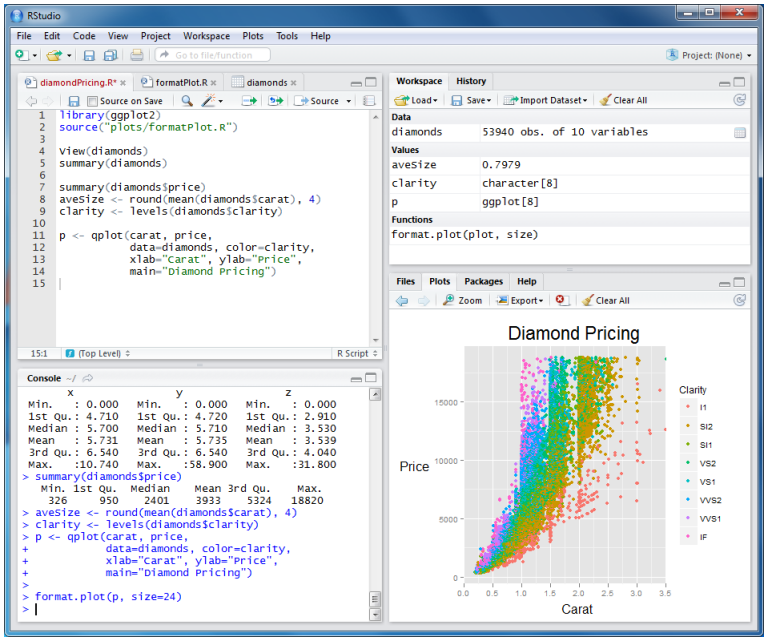
- Мане:
  - компликованост (у односу на друге статистичке пакете)
  - спорост, нарочито при раду са великим базама података
  - немогућност вршења симболичких израчунавања
  - ограниченост графичког корисничког окружења

- Коришћење R-а:
  - R console GUI
  - Rstudio (интерактивно развојно окружење)
  - Tinn-R editor и R console
  - Notepad++
  - …

- RStudio 

https://www.rstudio.com/

  - "*user-friendly*" окружење
  - четири прозора (одозго-надоле, слева-надесно, редом): едитор, конзола, историја, документи/графици/...
  - аутоматско допуњавање кода
  - истицање затворених заграда и наводника
  - инсталирање пакета из менија
  - позивање пакета из менија са стране
  - пречице на тастатури
  - лака претрага историје
  - ...



- Основе
  - Први кораци:
    - “прерасли”, интерактивни дигитрон
    - додела вредности објекту
    - векторска аритметика; стандардне процедуре; графици
  - Основе језика R:
    - изрази и објекти
    - функције и аргументи
    - вектори, матрице и вишедимензионалне матрице, фактори, листе, базе података
    - индексирање елемен(а)та објекта
    - условна селекција елемен(а)та објекта
    - сортирање

- R окружење
  - Управљање сесијом:
    - радни простор
    - текстуални излаз
    - R помоћ
    - пакети
    - уграђени подаци; `attach`/`detach`; `subset`, `transform` и `within`
  - Графички подсистем:
    - формат графика
    - “састављање” графика из делова
    - коришћење функције `par`
    - комбиновање графика
  - R програмирање:
    - контрола тока (условна извршења и петље)
    - класе и генеричке функције (`print`, `summary`)
  - Унос података:
    - учитавање података из текстуалног фајла; `read.table`, `read.csv2`...

- Вероватноћа и расподеле
  - Случајно узорковање (`sample`)
  - Израчунавање вероватноћа и комбинаторика
  - Уграђене расподеле у R-у:
    - густине расподела вероватноће
    - функције расподела вероватноће
    - квантили
    - случајни бројеви
- Статистика
  - Основне статистичке функције, сумарне статистике
  - Графички приказ расподела:
    - хистограми
    - емпиријска функција расподеле
    - *Q-Q plots*
    - *boxplots*
  - Табеле
  - Графички приказ табела
    - барови
    - питице

- Базе података (*data frame/matrix/set*)

  То је листа сачињена од вектора и/или фактора једнаке дужине који су у таквој међусобној вези да подаци на истој позицији у сваком од њих потичу од исте експерименталне јединице (субјекта, животиње и сл).

  Може се замислити и као матрица чије су врсте случајеви (опсервације), а колоне променљиве.

  Базе података могу се лако конструисати од (постојећих или нових) вектора коришћењем функције `data.frame()`.

```
> colors <- c("red","yellow","blue")
> numbers <- c(1, 2, 3)
> col.and.num<- data.frame(colors, numbers, more.numbers=4:6)
> col.and.num
  colors numbers more.numbers
1    red       1            4
2 yellow       2            5
3   blue       3            6
> str(col.and.num)
'data.frame':   3 obs. of  3 variables:
 $ colors      : Factor w/ 3 levels "blue","red","yellow": 2 3 1
 $ numbers     : num  1 2 3
 $ more.numbers: int  4 5 6
```

Обраћање подацима у бази, тзв. индексирање, и условна селекција:

```
> col.and.num[3,]
  colors numbers more.numbers
3   blue       3            6
> col.and.num[col.and.num$numbers<=2,]
  colors numbers more.numbers
1    red       1            4
2 yellow       2            5
> col.and.num[col.and.num$numbers<=2 & col.and.num$more.numbers>4,][1]
  colors
2 yellow
> col.and.num[,2]
[1] 1 2 3
> col.and.num[,"numbers"]
[1] 1 2 3
```

“Прилепљивање” нових колона, односно врста, на постојећу базу:

```
> exact <- c(T, F, F)
> col.and.num1 <- cbind(col.and.num, exact)
> col.and.num1
  colors numbers more.numbers exact
1    red       1            4  TRUE
2 yellow       2            5 FALSE
3   blue       3            6 FALSE
> zero <- c("red", 0, 0, F)
> col.and.num2 <- rbind(zero, col.and.num1) #PAZNJA: ne dobija se uvek zeljeni rezultat
> str(col.and.num2)
'data.frame':   4 obs. of  4 variables:
 $ colors      : Factor w/ 3 levels "blue","red","yellow": 2 2 3 1
 $ numbers     : chr  "0" "1" "2" "3"
 $ more.numbers: chr  "0" "4" "5" "6"
 $ exact       : chr  "FALSE" "TRUE" "FALSE" "FALSE"
```

Коришћење база података, које се налазе у оквиру R пакета:

```
> data(trees)
> d <- trees
> str(d)
'data.frame':  31 obs. of  3 variables:
 $ Girth : num  8.3 8.6 8.8 10.5 10.7 10.8 11 11 11.1 11.2 ...
 $ Height: num  70 65 63 72 81 83 66 75 80 75 ...
 $ Volume: num  10.3 10.3 10.2 16.4 18.8 19.7 15.6 18.2 22.6 19.9 ...
> (saved.names <- names(d))
[1] "Girth"  "Height" "Volume"
> (names(d) <- paste("Var", 1:dim(d)[2], sep="."))
[1] "Var.1" "Var.2" "Var.3"
> (names(d) <- saved.names)
[1] "Girth"  "Height" "Volume"
> str(d)
'data.frame':  31 obs. of  3 variables:
 $ Girth : num  8.3 8.6 8.8 10.5 10.7 10.8 11 11 11.1 11.2 ...
 $ Height: num  70 65 63 72 81 83 66 75 80 75 ...
 $ Volume: num  10.3 10.3 10.2 16.4 18.8 19.7 15.6 18.2 22.6 19.9 ...
> d$Height
 [1] 70 65 63 72 81 83 66 75 80 75 79 76 76 69 75 74 85 86 71 64 78 80 74 72 77 81 82
[28] 80 80 80 87
> d$Height[1:5]
[1] 70 65 63 72 81
> d[1,"Height"]
[1] 70
> d[1,]$Height
[1] 70
```

Сортирање базе података:

Када је потребно сортирати вектор најједноставније је то учинити позивом функције `sort()`. За сортирање базе података користи се функција `order()`, именовањем једног или више поља на основу којих би требало извршити сортирање.

```
> d[order(d$Height, d$Girth),]
   Girth Height Volume
3    8.8     63   10.2
20  13.8     64   24.9
2    8.6     65   10.3
…
4   10.5     72   16.4
24  16.0     72   38.3
16  12.9     74   22.2
23  14.5     74   36.3
…
18  13.3     86   27.4
31  20.6     87   77.0
```

Сортирање на основу већег броја критеријума врши се прослеђивањем вишеструких аргумената функцији `order()`. Друга променљива се користи када се о уређењу не може одлучити на основу прве променљиве.

Сортирање се по *default*-у врши у растућем поретку. Поредак сортирања може се променити, тј. може се вршити сортирање у опадајућем поретку, додавањем знака минус испред променљиве.

```
> d[order(d$Height, -d$Girth),]
   Girth Height Volume
3    8.8     63   10.2
20  13.8     64   24.9
2    8.6     65   10.3
…
24  16.0     72   38.3
4   10.5     72   16.4
23  14.5     74   36.3
16  12.9     74   22.2
…
18  13.3     86   27.4
31  20.6     87   77.0
```

Променљиве у бази података могу се учинити видљивим на глобалном нивоу у R изразима функцијом `attach()`. Овим се избегава гломазна $-нотација.

```
> attach(d)
> HG.ratio <- Height/Girth
> HG.ratio
 [1] 8.433735 7.558140 7.159091 6.857143 7.570093 7.685185 6.000000 6.818182 7.207207
[10] 6.696429 6.991150 6.666667 6.666667 5.897436 6.250000 5.736434 6.589147 6.466165
[19] 5.182482 4.637681 5.571429 5.633803 5.103448 4.500000 4.723926 4.682081 4.685714
[28] 4.469274 4.444444 4.444444 4.223301
> str(HG.ratio)
 num [1:31] 8.43 7.56 7.16 6.86 7.57 ...
> detach(d)
```

Оно што се, заправо, дешава приликом позива ове функције јесте да се одређена база података смешта у *system's search path*. Одатле се база података може уклонити функцијом `detach()`.

Мана коришћења функције `attach()` огледа се у томе што имена променљивих могу "сакрити" или бити "сакривена" другим објектима. R у понеким случајевима даје упозорење о томе, али не увек. Алтернатива је функција `with()`.

```
> with(trees, Height/Girth)
 [1] 8.433735 7.558140 7.159091 6.857143 7.570093 7.685185 6.000000 6.818182 7.207207
[10] 6.696429 6.991150 6.666667 6.666667 5.897436 6.250000 5.736434 6.589147 6.466165
[19] 5.182482 4.637681 5.571429 5.633803 5.103448 4.500000 4.723926 4.682081 4.685714
[28] 4.469274 4.444444 4.444444 4.223301
```

Издвајање подскупова података и креирање нових база са трансформисаним променљивим:

```
> (d.small <- subset(d, Volume < 18))
  Girth Height Volume
1   8.3     70   10.3
2   8.6     65   10.3
3   8.8     63   10.2
4  10.5     72   16.4
7  11.0     66   15.6
> (d.in_m3 <- transform(d, Volume.m=0.0283168466*Volume))
   Girth Height Volume  Volume.m
1    8.3     70   10.3 0.2916635
2    8.6     65   10.3 0.2916635
3    8.8     63   10.2 0.2888318
4   10.5     72   16.4 0.4643963
…
29  18.0     80   51.5 1.4583176
30  18.0     80   51.0 1.4441592
31  20.6     87   77.0 2.1803972
```

Алтернатива `transform()` је функција `within()`.

```
> (d.in_m3_c <- within(d, {
+   Volume.m <- 0.0283168466*Volume
+   m <- mean(Volume.m)
+   centered.Volume.m <- Volume.m - m
+   rm(m)
+ }))
   Girth Height Volume centered.Volume.m  Volume.m
1    8.3     70   10.3       -0.56268315 0.2916635
2    8.6     65   10.3       -0.56268315 0.2916635
3    8.8     63   10.2       -0.56551483 0.2888318
4   10.5     72   16.4       -0.38995038 0.4643963
…
29  18.0     80   51.5        0.60397093 1.4583176
30  18.0     80   51.0        0.58981251 1.4441592
31  20.6     87   77.0        1.32605052 2.1803972
```

Често се дешава да су подаци добијени од вишеструких извора, па је потребно сјединити их у већу целину.

Вертикално “слагање”:

```
> data(iris)
> #simulacija podataka dobijenih od visestrukih izvora
> se <- subset(iris, Species=="setosa", select=-c(Species, Sepal.Length))
> ve <- subset(iris, Species=="versicolor", select=-c(Species, Petal.Width))
> vi <- subset(iris, Species=="virginica", select=-c(Species))
> s.length.se <- iris[[1]][iris$Species=="setosa"]
> p.width.ve <- iris[[4]][iris$Species=="versicolor"]
> #spajanje
> se$Species <- factor("Iris setosa")
> se$Sepal.Length <- s.length.se
> ve$Species <- factor("Iris versicolor")
> ve$Petal.Width <- p.width.ve
> vi$Species <- factor("Iris virginica")
> irisall <- rbind(se, ve, vi)
> names(irisall)
[1] "Sepal.Width"  "Petal.Length" "Petal.Width"  "Species"      "Sepal.Length"
> levels(irisall$Species)
[1] "Iris setosa"     "Iris versicolor" "Iris virginica"
```

Захтева се само да базе које се спајају садрже исте променљиве, иако није неопходно да оне буду у истом поретку у свакој од база.

Слично ситуацији када су подаци за различите групе субјеката организовани у више база, могућа је и ситуација да су различите врсте података за исте субјекте засебно прикупљене.

Није препоручљиво коришћење функције `cbind()`.

```
> #dve baze spajaju se po promenljivoj, istog imena (u obe baze)
> #ta promenljiva je obicno ID - cuva identifikaciju subjekta
> data1 <- data.frame(ID=1:5, x=letters[1:5])
> data2 <- data.frame(ID=1:5, y=letters[6:10])
> merge(data1, data2)
  ID x y
1  1 a f
2  2 b g
3  3 c h
4  4 d i
5  5 e j

> #slucaj kada postoji vise promenljivih istog imena (u obe baze)
> data1 <- data.frame(ID=1:5, x=letters[1:5])
> data2 <- data.frame(ID=1:5, x=letters[6:10])
> merge(data1, data2, by="ID")
  ID x.x x.y
1  1   a   f
2  2   b   g
3  3   c   h
4  4   d   i
5  5   e   j
> merge(data1, data2, by="ID", suffixes=c(1, 2))
  ID x1 x2
1  1  a  f
2  2  b  g
3  3  c  h
4  4  d  i
5  5  e  j
```

```
> #po default-u argument 'all' postavljen je na vrednost 'F'
> #kao rezultat vracaju se samo vrste iz baza, koje su odgovarajuce
> merge(data1, data2)
[1] ID x
<0 rows> (or 0-length row.names)
> merge(data1, data2, all=TRUE)
   ID x
1   1 a
2   1 f
3   2 b
4   2 g
5   3 c
6   3 h
7   4 d
8   4 i
9   5 e
10  5 j
> data1 <- data.frame(ID=1:5, x=letters[1:5])
> data2 <- data.frame(ID=4:8, y=letters[6:10])
> merge(data1, data2)
  ID x y
1  4 d f
2  5 e g
> merge(data1, data2, all=TRUE)
  ID    x    y
1  1    a <NA>
2  2    b <NA>
3  3    c <NA>
4  4    d    f
5  5    e    g
6  6 <NA>    h
7  7 <NA>    i
8  8 <NA>    j
> merge(data1, data2, all.x=TRUE)
  ID x    y
1  1 a <NA>
2  2 b <NA>
3  3 c <NA>
4  4 d    f
5  5 e    g
```

```
> #baze nickel i ewrates nalaze se u okviru paketa ISwR
> head(nickel)
  id icd exposure      dob  age1st   agein  ageout
1  3   0        5 1889.019 17.4808 45.2273 92.9808
2  4 162        5 1885.978 23.1864 48.2684 63.2712
3  6 163       10 1881.255 25.2452 52.9917 54.1644
4  8 527        9 1886.340 24.7206 47.9067 69.6794
5  9 150        0 1879.500 29.9575 54.7465 76.8442
6 10 163        2 1889.915 21.2877 44.3314 62.5413
> head(ewrates)
  year age lung nasal other
1 1931  10    1     0  1269
2 1931  15    2     0  2201
3 1931  20    6     0  3116
4 1931  25   14     0  3024
5 1931  30   30     1  3188
6 1931  35   68     1  4165
> nickel <- transform(nickel,
+ agr = trunc(agein/5)*5,
+ ygr = trunc((dob+agein-1)/5)*5+1)
> mrg <- merge(nickel, ewrates, by.x=c("agr","ygr"), by.y=c("age","year"))
> head(mrg,10)
   agr  ygr  id icd exposure      dob  age1st   agein  ageout lung nasal other
1   20 1931 273 154        0 1909.500 14.6913 24.7465 55.9302    6     0  3116
2   20 1931 213 162        0 1910.129 14.2018 24.1177 63.0493    6     0  3116
3   20 1931 546   0        0 1909.500 14.4945 24.7465 72.5000    6     0  3116
4   20 1931 574 491        0 1909.729 14.0356 24.5177 70.6592    6     0  3116
5   20 1931 110   0        0 1909.247 14.0302 24.9999 72.7534    6     0  3116
6   20 1931 325 434        0 1910.500 14.0737 23.7465 43.0343    6     0  3116
7   25 1931  56 502        2 1904.500 18.2917 29.7465 51.5847   14     0  3024
8   25 1931 690 420        0 1906.500 17.2206 27.7465 55.1219   14     0  3024
9   25 1931 443 420        0 1905.326 14.5562 28.9204 65.7616   14     0  3024
10  25 1931 137 465        0 1905.386 19.0808 28.8601 74.2794   14     0  3024
```

Спајање ових база, од којих прва описује еснаф топионичара никла у Велсу, а друга садржи таблицу смртности по годинама и старосним групама у петогодишњим интервалима, извршено је на основу вредности приликом приступања субјекта истраживању.

Увоз података:

```
> #working directory
> getwd() #provera koji je folder radni
[1] "C:/Users/Lenchy/Documents"
> setwd("C:/moj_projekat") #promena radnog foldera
```

Табеларни подаци у текстуалном документу:

```
> my.data <- read.table("CROATRAD.txt", header=T)
> str(my.data)
'data.frame':  125 obs. of  7 variables:
 $ X40K   : num  13.7 14 12.3 13.2 11.3 15.8 14.3 15.5 13.3 12.6 ...
 $ X224Ra : num  41.9 52.6 40.6 45.9 40.6 20 45.3 50.6 59.2 37.9 ...
 $ X228Ra : num  24.4 48.4 27.7 29.6 21.6 8.9 25.4 28.2 30.6 41.8 ...
 $ X236U  : num  0.1 0.4 0.2 0.4 0.1 0.1 0.2 0.2 0.2 0.2 ...
 $ X137Cs : num  2.8 3.3 1.6 2.5 1.6 6 1.1 1.7 1.8 2.4 ...
 $ Natural: num  80.2 115.4 80.8 89.2 73.6 ...
 $ Total  : num  83 118.7 82.3 91.6 75.2 ...
```

Захтева се да подаци буду у ASCII формату, садржани у једноставном документу, без икаквог форматирања, креираном нпр. у Notepad-у или сличним текстуалним едиторима.

```
> my.data1 <- read.table("C:/Users/Lenchy/Desktop/posao_fax/TU/CROATRAD.txt", header=T)
> str(my.data1)
'data.frame':  125 obs. of  7 variables:
 $ X40K   : num  13.7 14 12.3 13.2 11.3 15.8 14.3 15.5 13.3 12.6 ...
 $ X224Ra : num  41.9 52.6 40.6 45.9 40.6 20 45.3 50.6 59.2 37.9 ...
 $ X228Ra : num  24.4 48.4 27.7 29.6 21.6 8.9 25.4 28.2 30.6 41.8 ...
 $ X236U  : num  0.1 0.4 0.2 0.4 0.1 0.1 0.2 0.2 0.2 0.2 ...
 $ X137Cs : num  2.8 3.3 1.6 2.5 1.6 6 1.1 1.7 1.8 2.4 ...
 $ Natural: num  80.2 115.4 80.8 89.2 73.6 ...
 $ Total  : num  83 118.7 82.3 91.6 75.2 ...
```

Подаци у документу CSV (Comma Delimited Values) формата:

```
> my.data2 <- read.csv2("C:/Users/Lenchy/Desktop/posao_fax/TU/CROATRADc.csv", header=T)
> str(my.data2)
'data.frame':  125 obs. of  7 variables:
 $ X40K   : num  13.7 14 12.3 13.2 11.3 15.8 14.3 15.5 13.3 12.6 ...
 $ X224Ra : num  41.9 52.6 40.6 45.9 40.6 20 45.3 50.6 59.2 37.9 ...
 $ X228Ra : num  24.4 48.4 27.7 29.6 21.6 8.9 25.4 28.2 30.6 41.8 ...
 $ X236U  : num  0.1 0.4 0.2 0.4 0.1 0.1 0.2 0.2 0.2 0.2 ...
 $ X137Cs : num  2.8 3.3 1.6 2.5 1.6 6 1.1 1.7 1.8 2.4 ...
 $ Natural: num  80.2 115.4 80.8 89.2 73.6 ...
 $ Total  : num  83 118.7 82.3 91.6 75.2 ...
```

Подаци у Excel-у:

```
> #ucita se paket gdata
> my.data3 <- read.xls("C:/Users/Lenchy/Desktop/posao_fax/TU/CROATRADee.xls", header=T)
> str(my.data3)
'data.frame':  125 obs. of  7 variables:
 $ X40K   : num  13.7 14 12.3 13.2 11.3 15.8 14.3 15.5 13.3 12.6 ...
 $ X224Ra : num  41.9 52.6 40.6 45.9 40.6 20 45.3 50.6 59.2 37.9 ...
 $ X228Ra : num  24.4 48.4 27.7 29.6 21.6 8.9 25.4 28.2 30.6 41.8 ...
 $ X236U  : num  0.1 0.4 0.2 0.4 0.1 0.1 0.2 0.2 0.2 0.2 ...
 $ X137Cs : num  2.8 3.3 1.6 2.5 1.6 6 1.1 1.7 1.8 2.4 ...
 $ Natural: num  80.2 115.4 80.8 89.2 73.6 ...
 $ Total  : num  83 118.7 82.3 91.6 75.2 ...
```

База података коришћена у претходна два слајда садржи 125 мерења из ваздуха радијације емитоване од стране ^{137}Cs и укупне радијације, забележена у Истри (Хрватска).

Цезијум 137 је вештачки радионуклид, који настаје фисијом. Није постојао у природи пре почетка нуклеарних проба и несрећа на нуклеарним постројењима (Чернобил).

Извор: *“Statistics and Data Analysis in Geology”*, J. C. Davis