

Увод у R

1. час

7. март 2016.

- 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

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. 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

“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

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

```
> data(trees)
> d <- trees
> head(d)
```

	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
5	10.7	81	18.8
6	10.8	83	19.7

```
> d$Height
[1] 70 65 63 72 81 83 66 75 80 75 79 76 76 69 75 74 85 86 71 64 78 80 74 72 77 81 82 80 80 80 87
> d$Height[1:10]
[1] 70 65 63 72 81 83 66 75 80 75
> d$Height[1]
[1] 70
```

Сортирање базе података

За сортирање вектора најједноставније је користити функцију `sort()`. За сортирање базе података користи се функција `order()`, именовањем једног или више поља на основу којих би требало извршити сортирање. 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.

```
> d[order(d$Height, d$Girth),]
```

	Girth	Height	Volume
3	8.8	63	10.2
20	13.8	64	24.9
...			
16	12.9	74	22.2
23	14.5	74	36.3
8	11.0	75	18.2
10	11.2	75	19.9
...			
18	13.3	86	27.4
31	20.6	87	77.0

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.

Учитавање података

Учитавање табеле као базе података врши се помоћу функције `read.table("putanja/ime_fajla.txt",header=T)`.

Аргумент `header=T` означава да прву врсту не посматра као податке већ као `header`. На пример, прво правимо текстуални фајл `cars.txt` који сместимо у исти фолдер или у позиву функције `read.table` у имену фајла наводимо целу путању.

```
baza<-read.table("cars.txt",header=T)
```

Ова функција подразумева да је сепаратор празно поље.

Могуће је да сепаратор буде и

- зарез, тада се користи функција `read.csv()`
- ТАВ, тада се користи функција `read.delim()`

Уколико се подаци налазе у *excel*-у, могуће их је учитати помоћу:

```
baza <- read.xls("putanja/ime_fajla.txt", header=T)
```

Табеле

Код експерименталних података често се појављују категоричке променљиве, које указују на неку потподелу података у групе. Њих у R-у означавамо као факторе (factors). Исправност многих анализа у R-у зависи од коректног препознавања фактора, нпр. функција `table()` ради само са категоричким променљивим. Фактори се дефинишу функцијом `factor()`, односно `ordered()`. Могу се добити и конверзијом знаковних или нумеричких вектора функцијама за промену типа: `as.factor()` и `as.ordered()`. Нивои постојећег фактора издвајају се функцијом `levels()`.

```
> table(trees$Height)
```

```
63 64 65 66 69 70 71 72 74 75 76 77 78 79 80 81 82 83 85 86 87
 1  1  1  1  1  1  1  2  2  3  2  1  1  1  5  2  1  1  1  1  1
```

Иако сличне, табеле нису у потпуности еквивалентне матрицама. Постоји класа објеката "table" за коју постоје специфични методи. Важно је поменути један опционални аргумент функције `table()`, а то је `useNA`, који може бити постављен на једну од вредности "no", "ifany", "always".

Графички приказ табела

Графичке могућности су важна и изузетно разноврсна компонента R-овог окружења. Обезбеђене су функције за генерисање различитих статистичких графика од података, као и флексибилно окружење за модификовање детаља на графицима. Могу се “градити” и потпуно нове врсте графика, прилагођене конкретном проблему. Брз увид у обиље R-ових графика стиче се покретањем програма за демонстрацију:

```
> demo(graphics)
> demo(image)
> demo(lattice)
```

Наредбе за креирање графова деле се у три основне групе:

- функције за графике високог нивоа креирају нови графикон, могуће са осама, ознакама, насловима итд.
- функције за графике ниског нивоа додају више информација на постојећи графикон, нпр. додатне тачке, линије и ознаке
- функције за интерактивне графике омогућавају да се интерактивно дода информација на постојећи графикон или из њега “извуче” информација, нпр. помоћу компјутерског миша

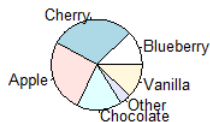
Питице (the pie chart)

```
> pie.sales
```

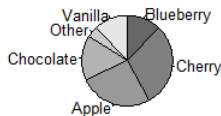
Blueberry	Cherry	Apple	Chocolate	Other	Vanilla
0.12	0.30	0.26	0.16	0.04	0.12

```
> pie(pie.sales, main = "Obicna pitica")  
> pie(pie.sales, col = gray(seq(0.4, 0.9, length = 6)), clockwise = TRUE, main = "Nijanse sive")  
> pie(pie.sales, col = rainbow(6), clockwise = TRUE, main = "Boje duge")  
> library(plotrix)  
> pie3D(pie.sales, explode = 0.1, main = "3D pitica", labelcex = 0.6)
```

Obicna pitica



Nijanse sive



Boje duge



3D pitica

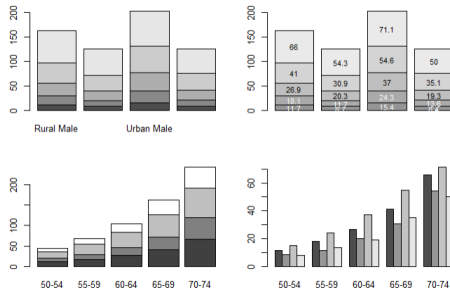


Барови (the bar chart)

```
> VADeaths
```

	Rural Male	Rural Female	Urban Male	Urban Female
50-54	11.7	8.7	15.4	8.4
55-59	18.1	11.7	24.3	13.6
60-64	26.9	20.3	37.0	19.3
65-69	41.0	30.9	54.6	35.1
70-74	66.0	54.3	71.1	50.0

```
> barplot(VADeaths)
> barplot(VADeaths, col=gray(0.5 + 1:5/12), names=rep("", 4))
> text(rep(midpts, each=5), apply(VADeaths, 2, cumsum) - VADeaths/2, VADeaths, col=rep(c("white", "black"),
+ times=2:3), cex=0.8)
> barplot(t(VADeaths), col = gray(c(0.25,0.5,0.75, 1)))
> barplot(t(VADeaths), beside = TRUE)
```



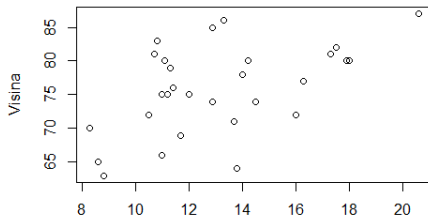
Функција `plot()`

Једна је од R-ових најчешће коришћених функција за графике; спада међу генеричке функције, тј. врста графика који креира дијаграм у зависности од врсте/класе првог аргумента.

- **`plot(x, y)`, `plot(xy)`**

Ако су x и y вектори, `plot(x, y)` за резултат има тачкасти график (`scatterplot`) y -а у зависности од x -а. Исти учинак може се произвести прослеђивањем само једног аргумента xy функцији `plot()`, који је, у том случају, листа са два вектора или матрица са две колоне.

> `plot(cbind(trees$Girth, trees$Height), xlab="Obim", ylab="Visina")`

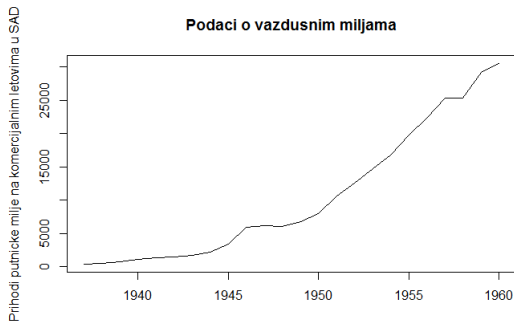


Функција plot()

- **plot(x)**

Ако је x временска серија (time series), резултат је граф временске серије; ако је x нумерички вектор на излазу се добија график вредности тог вектора у зависности од индекса у вектору.

```
> plot(airmiles, main="Podaci o vazдушnim miljama", xlab="vreme", ylab="Prihodi putnicke  
milje na komercijalnim letovima u SAD")
```



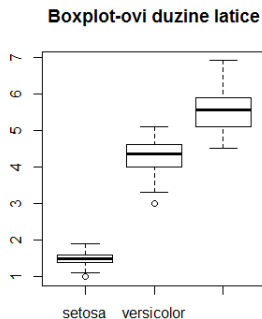
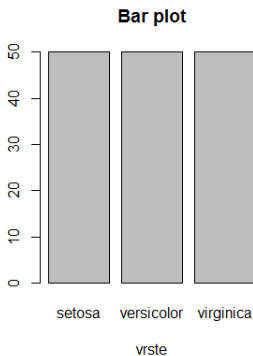
Функција plot()

- **plot(f), plot(f, y)**

Ако је *f* је фактор, *y* је нумерички вектор; првом наредбом генерише се график *f*-а у облику bar-ова, док се као резултат друге наредбе добијају boxplot-ови *y*-а за сваки ниво *f*-а.

> plot(iris\$Species, xlab="vrste", main="Bar plot")

> plot(iris\$Species, iris\$Petal.Length, main="Boxplot-ovi duzine latice")

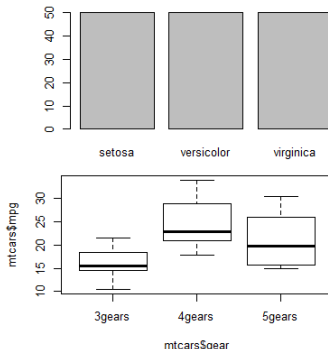
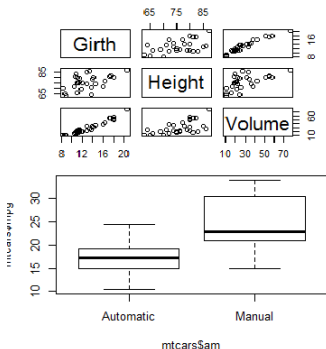


Функција plot()

- **plot(df), plot(~ expr), plot(y ~ expr)**

df је база података, y је било који објекат, expr је списак назива објеката одвојених помоћу '+'; првим двама наредбама, редом, генеришу се distribution plots променљивих у бази, односно distribution plots броја именованих објеката; излаз треће наредбе је низ графика y-а у односу на сваки објекат именован у expr.

```
> plot(trees)
> plot(~ iris$Species)
> plot(mtcars$mpg ~ mtcars$am + mtcars$gear)
```



У R постоје и две веома корисне функције за презентовање вишеструких података.

- **pairs(X)**

где је X нумеричка матрица или база података. Овом наредбом се добија матрица тачкастих графика парова променљивих дефинисаних колонама X -а, тј. свака колона X -а приказана је у зависности од сваке друге колоне X -а.

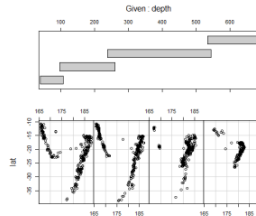
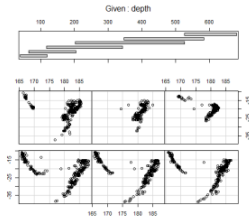
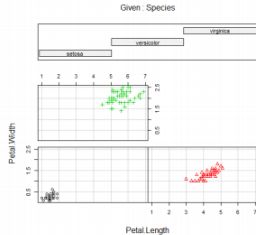
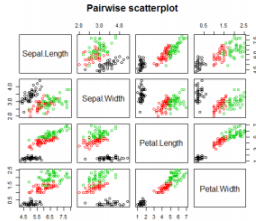
- **coplot(a ~ b | c)**

где су a и b нумерички вектори, а c вектор или фактор (претпоставља се да су сви једнаке дужине). Излаз је низ графика на којима је приказана зависност a од b за дате вредности c -а. Ако је c фактор, то значи да је a приказан у односу на b за сваки ниво c -а. Ако је, пак, c нумерички вектор, он је издељен на неколико “условних” интервала и за сваки интервал a је приказан у односу на b за вредности c -а унутар тог интервала.

```

> pairs(iris[,1:4], col=as.numeric(Species), main="Pairwise scatterplot")
> coplot(Petal.Width ~ Petal.Length | Species, col=as.numeric(Species), pch=as.nume
> coplot(lat ~ long | depth, data = quakes)
> given.depth <- co.intervals(quakes$depth, number = 4, overlap = .1)
> coplot(lat ~ long | depth, data = quakes, given.v = given.depth, rows = 1)

```

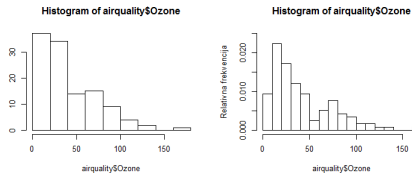


Хистограм

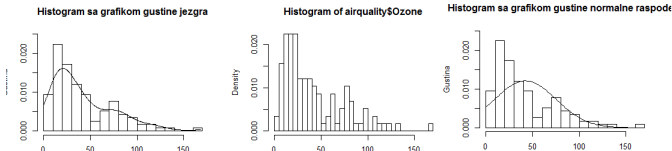
Хистограм је најједноставнија непараметарска “оцена” густине расподеле. Функција **hist(x)**, где је x нумерички вектор је наредба за исцртавање хистограма. То је посебна врста графика са bar-овима, која се користи за приказ расподеле фреквенција низа (непрекидних) нумеричких вредности. Од опционалних аргумената важно је издвојити:

- `nclass`
постављањем на одређену вредност корисник може да препоручи жељени (разуман) број bar-ова
- `breaks` (`breaks= "Sturges"` / `breaks="Scott"` / `breaks="Freedman-Diaconis"`)
могу се тачно задати број bar-ова, односно граничне тачке bar-ова
- `prob=T`
вредности на у-оси представљају релативне, а не апсолутне фреквенције

```
> hist(airquality$Ozone)
> hist(airquality$Ozone, breaks=15, prob=T, ylim=c(0, 0.025), ylab="Relativna frekvencija")
```



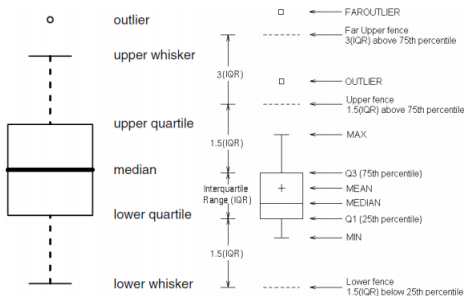
```
> max.ozone <- max(airquality$Ozone, na.rm=T)
> hist(airquality$Ozone, prob=T, breaks=15, xlab='Ozone (ppb)', ylim=c(0, 0.025),
+ ylab='Gustina', main='Histogram sa grafikom gustine jezgra')
> lines(density(airquality$Ozone, na.rm=T, from=0, to=max.ozone))
> mean.ozone <- mean(airquality$Ozone, na.rm=T)
> sd.ozone <- sd(airquality$Ozone, na.rm=T)
> ozone.histogram <- hist(airquality$Ozone, breaks=50, prob=T)
> hist(airquality$Ozone, breaks=15, prob=T, ylim=c(0, 0.025), xlab='Ozone (ppb)',
+ ylab='Gustina', main='Histogram sa grafikom gustine normalne raspodele')
> curve(dnorm(x, mean=mean.ozone, sd=sd.ozone), add = T)
```



Boxplot

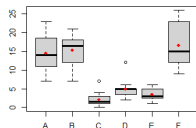
Функција **boxplot(x)**, где је x нумерички вектор је наредба за исцртавање правоугаоног дијаграма. Приликом статистичке анализе требало би га обавезно користити за приказивање (непрекидне) нумеричке променљиве, чије су вредности измерене на подгрупама субјеката, одређеним категоричком променљивом.

Висина правоугаоника (боха) дефинисана је распоном између квантила (IQR). The lower whisker (доња "ограда") повлачи се са доње стране боха и налази се на нивоу који одговара већој од две вредности: узорачки минимум и први квантил умањен за 1.5IQR . Слично, the upper whisker (горња "ограда") повлачи се са горње стране боха и налази се на нивоу који одговара мањој од две вредности: узорачки максимум и трећи квантил увећан за 1.5IQR .



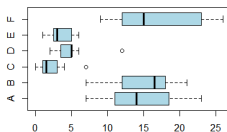
Boxplot

```
> boxplot(count ~ spray, data = InsectSprays, col = "lightgray")  
> means <- tapply(InsectSprays$count, InsectSprays$spray, mean)  
> points(means, col="red", pch=18)
```



Уместо вертикалне оријентације boxplot-ова, навођењем опционалног аргумента `horizontal=T`, добијају се хоризонтално постављени boxplot-ови.

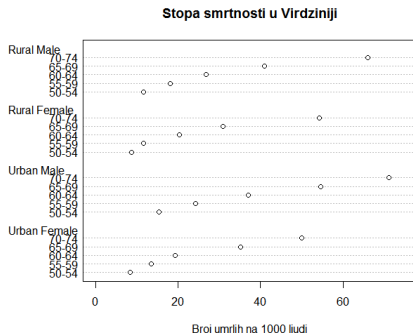
```
> boxplot(count ~ spray, data = InsectSprays, horizontal=T, col = "lightgray")
```



Dotchart

Функција **dotchart()** је функција за генерисање тачкастог графика. Посебно је погодан за визуални приказ свих унесених података с вредностима које леже у одређеним распонима.

```
> dotchart(VADeaths, xlim=c(0,75), xlab="Broj umrlih na 1000 ljudi",  
           main="Stopa smrtnosti u Virdziniji")
```



Q-Q графици

Q-Q графици су врста тачкастих графика, који се користе да би се упоредиле расподеле две групе података или узорачка са (неком) референтном расподелом.

- **qqnorm(x)**
исцртава нумерички вектор x у односу на очекивану нормалну расподелу
- **qqline(x)**
додаје праву на претходни график која пролази кроз квантиле расподеле и података
- **qqplot(x, y)**
исцртава квантиле x -а у односу на квантиле y -а ради поређења њихових припадних расподела

Q-Q графици

```
> X <- rnorm(1000); A <- rnorm(1000); B <- rnorm(1000, mean=3, sd=2)
> C <- rt(1000, df=2); D <- exp(rnorm(1000))
> par(mfrow=c(2, 2), mar=c(3, 3, 2, 1))
> qqplot(X, A, main="A i X su iz iste raspodele")
> qqplot(X, B, main="B je reskaliran X")
> qqplot(X, C, main="C ima tezi rep")
> qqplot(X, D, main="D je pomerena u desno")
```

