ТЕОРИЈА УЗОРАКА час 4

26. март ’14.

- Статистичко моделирање (*statistical modelling*)

  Најтежи део статистичког посла јесте сам почетак и то зато што би требало одабрати прави приступ статистичкој анализи. Приступ зависи, пре свега, од природе података којима се располаже, а с тим у вези и са посебним питањем, на које би требало пронаћи адекватан одговор. Наиме, кључно је разумевање каква је променљива која се добија као "одговор" (*response variable*), и, са друге стране, каква је природа променљиве која је објашњава (*explanatory variable*). Променљива "одговор" није ништа друго до променљива чије варирање би требало сагледати и схватити, и, приликом визуализације, њене вредности се уцртавају на у-оси. Вредности променљиве "објашњења" бележе се на х-оси; оно на чему се настоји јесте да се одреди у којој је мери варирање променљиве "одговора" повезано са варирањем променљиве "објашњења". При томе, мора се узети у обзир и то на који начин променљиве у анализи "мере" вредности обележја (обележје - варијабилна квалитативна или квантитативна особина) којима су придружене.

Након идентификовања променљивих које се појављују у анализи може се, на релативно једноставан начин, приступити избору пригодног статистичког метода:

променљиве "објашњења"	
све променљиве "објашњења" су непрекидне нумеричке променљиве	**регресија (*regression*)**
све променљиве "објашњења" су категоричке променљиве	**дисперзиона анализа (ANOVA)**
променљиве "објашњења" су и нумеричке и категоричке променљиве	**анализа коваријације (ANCOVA)**

променљива "одговора"	
непрекидна	(нормална) регресија, ANOVA, ANCOVA
релативне фреквенције	логистичка регресија
апсолутне фреквенције	log-линеарни модели
бинарна	бинарна логистичка анализа
време смрти	анализа преживљавања

Циљ је, затим, одредити вредности параметара за одређени модел, што ће за резултат имати најбоље могуће пристајање модела подацима (*the best fit of the model to the data*).

- Регресија (*regression*)

  Нека су X и Y два дата обележја, тј. две случајне величине. Случајна величина $E(Y|X) = R(X)$ назива се регресија. Може се показати да је то функција која најбоље описује зависност Y од X у смислу да је средње квадратно одступање $E(Y - h(X))^2$ најмање ако је $h(X) = R(X)$

  Дакле, ако се претпостави да постоји функционална зависност између два обележја, онда је, у смислу минимума средње квадратног одступања, најбоља функција којом се описује зависност та два обележја управо регресија. Међутим, да би се могла одредити функција регресије $R(X)$ неопходно је познавање заједничке расподеле случајних величина X и Y .

  Ако заједничка расподела није позната, могуће је, на основу података из узорка одредити облик функционалне зависности Y од X и параметре изабране функције.

Облици зависности:

- линеарна $Y = aX + b$
- полиномијална $Y = a_n X^n + a_{n-1} X^{n-1} + \cdots + a_0$
- степена $Y = aX^b$
- експоненцијална $Y = ae^{bX}$
- $Y = \frac{1}{aX + b}$
- ...

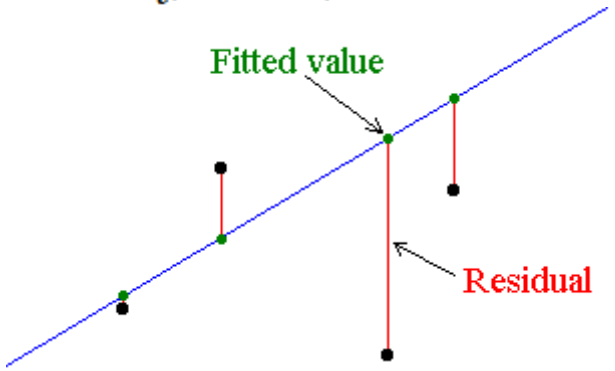
- Линеарна регресија (*simple linear regression*)

Линеарна зависност $Y = aX + b$ величина X и Y је најједноставнији облик зависности.

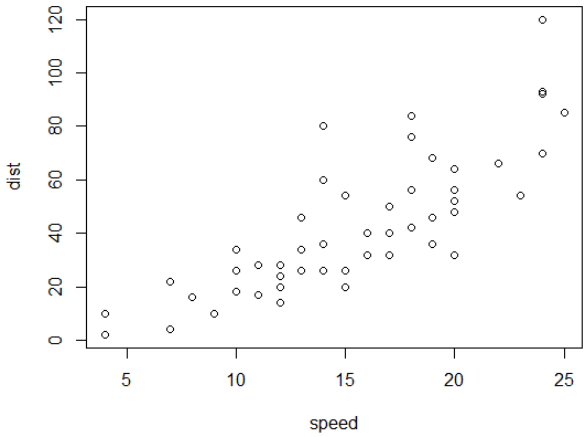
Ако, на основу података из узорка $(x_i, y_i), i = \overline{1,n}$, постоје назнаке да би се требало определити за линеарну зависност $f(x, a, b) = ax + b$, a и b се могу одредити из услова:

$$\min_{a,b \in \mathbb{R}} S(a,b) = \min_{a,b \in \mathbb{R}} \sum_{j=1}^{n} \left(y_j - (ax_j + b) \right)^2$$

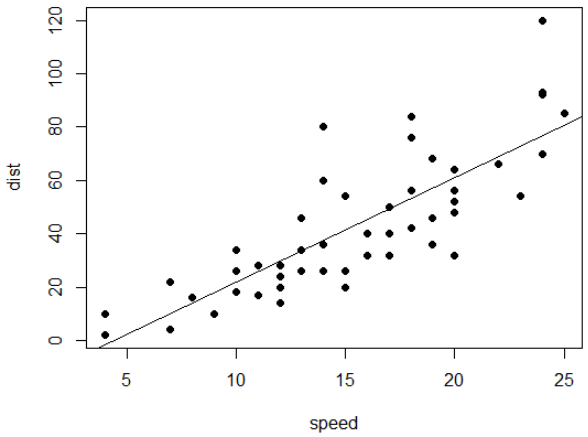
Геометријски, уствари, од свих правих $y = ax + b$ бира се она за коју је збир квадрата одсечака $d_i = y_i - (ax_i + b), i = \overline{1,n}$ најмањи. Величине $d_i, i = \overline{1,n}$ се називају <u>резидуали</u>.

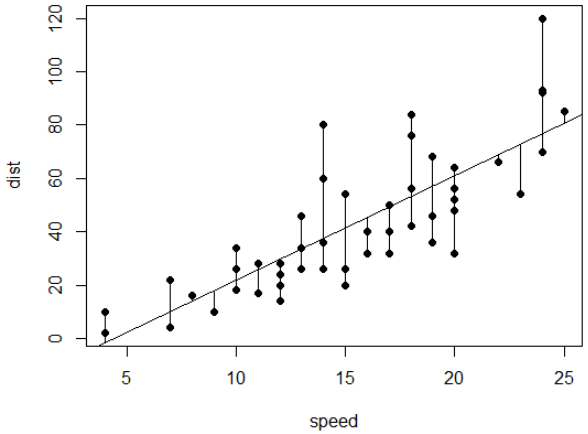



```
> attach(cars)
> cor(speed, dist)
[1] 0.8068949
> cor.test(speed, dist)

        Pearson's product-moment correlation

data:  speed and dist
t = 9.464, df = 48, p-value = 1.49e-12
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6816422 0.8862036
sample estimates:
      cor
0.8068949

> plot(dist ~ speed)
```

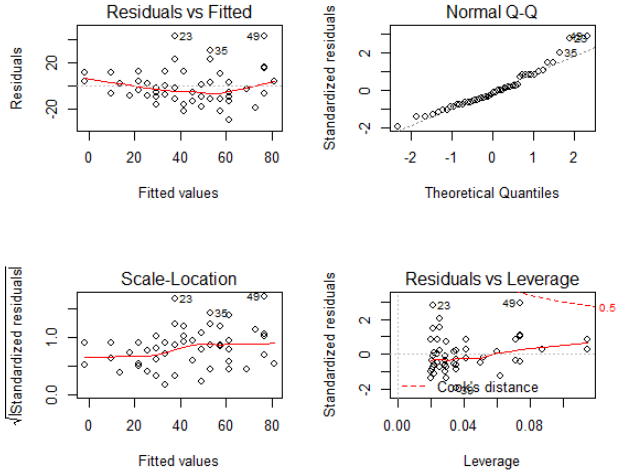


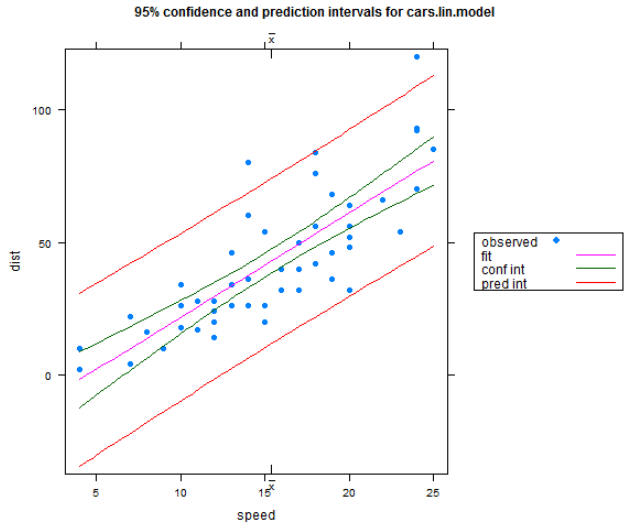
```
> cars.lin.model <- lm(dist ~ speed)
> cars.lin.model

Call:
lm(formula = dist ~ speed)

Coefficients:
(Intercept)        speed
     -17.58         3.93

> coef(cars.lin.model)
(Intercept)       speed
    -17.579       3.932
> plot(dist ~ speed, pch=16); abline(cars.lin.model)
#isti izlaz dobija se i naredbom: abline(coef(cars.lin.model))
> options(digits=4)
> summary(cars.lin.model)

Call:
lm(formula = dist ~ speed)

Residuals:
   Min     1Q Median     3Q    Max
-29.07  -9.53  -2.27   9.21  43.20

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -17.579      6.758   -2.60    0.012 *
speed          3.932      0.416    9.46  1.5e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.4 on 48 degrees of freedom
Multiple R-squared:  0.651,    Adjusted R-squared:  0.644
F-statistic: 89.6 on 1 and 48 DF,  p-value: 1.49e-12

> segments(speed, fitted(cars.lin.model), speed, dist)
```





```
> #objekat klase lm je lista imenovanih komponenti
> names(cars.lin.model)
 [1] "coefficients"  "residuals"     "effects"       "rank"          "fitted.values"
 [6] "assign"        "qr"            "df.residual"   "xlevels"       "call"
[11] "terms"         "model"
> #dostupne su razne f-je kojima se mogu preuzeti potrebne info. iz ove liste; primer su:
> fitted(cars.lin.model)
     1      2      3      4      5      6      7      8      9     10     11     12
-1.849 -1.849  9.948  9.948 13.880 17.813 21.745 21.745 21.745 25.677 25.677 29.610
    13     14     15     16     17     18     19     20     21     22     23     24
29.610 29.610 29.610 33.542 33.542 33.542 33.542 37.475 37.475 37.475 37.475 41.407
    25     26     27     28     29     30     31     32     33     34     35     36
41.407 41.407 45.339 45.339 49.272 49.272 49.272 53.204 53.204 53.204 53.204 57.137
    37     38     39     40     41     42     43     44     45     46     47     48
57.137 57.137 61.069 61.069 61.069 61.069 61.069 68.934 72.866 76.799 76.799 76.799
    49     50
76.799 80.731
> resid(cars.lin.model)
       1        2        3        4        5        6        7        8        9
  3.8495  11.8495  -5.9478  12.0522   2.1198  -7.8126  -3.7450   4.2550  12.2550
      10       11       12       13       14       15       16       17       18
 -8.6774   2.3226 -15.6098  -9.6098  -5.6098  -1.6098  -7.5422   0.4578   0.4578
      19       20       21       22       23       24       25       26       27
 12.4578 -11.4746  -1.4746  22.5254  42.5254 -21.4070 -15.4070  12.5930 -13.3394
      28       29       30       31       32       33       34       35       36
 -5.3394 -17.2719  -9.2719   0.7281 -11.2043   2.7957  22.7957  30.7957 -21.1367
      37       38       39       40       41       42       43       44       45
-11.1367  10.8633 -29.0691 -13.0691  -9.0691  -5.0691   2.9309  -2.9339 -18.8663
      46       47       48       49       50
 -6.7987  15.2013  16.2013  43.2013   4.2689
> #najcesce koriscena f-ja za razmatranje regresije je summary()
> #postoji plot metod namenjen bas lm objektima, koji daje "dijagnosticke" info.
> par(mfrow=c(2,2))
> plot(cars.lin.model)
```






```
> fitted(cars.lin.model)[1:5]
     1      2      3      4      5
-1.849 -1.849  9.948  9.948 13.880
> #predvidjanje
> pred.frame <- data.frame(speed=c(6, 8, 21))
> predict(cars.lin.model, newdata=pred.frame)
     1      2      3
 6.015 13.880 65.001
> predict(cars.lin.model, newdata=pred.frame, int="confidence")
     fit    lwr   upr
1  6.015 -2.973 15.00
2 13.880  6.308 21.45
3 65.001 58.597 71.41
> predict(cars.lin.model, newdata=pred.frame, int="prediction")
     fit    lwr   upr
1  6.015 -26.19 38.22
2 13.880 -17.96 45.72
3 65.001  33.42 96.58
> library(HH)
> ci.plot(cars.lin.model)
```

И у случају када је зависност обележја Y од X облика $Y = aX + b$ могу се запазити одступања података (мерених величина) $(x_i, y_i), i = \overline{1,n}$ од праве $y = ax + b$. Та одступања резултат су случајних грешака које се јављају приликом мерења. То значи да се узима следећи модел зависности Y од X:

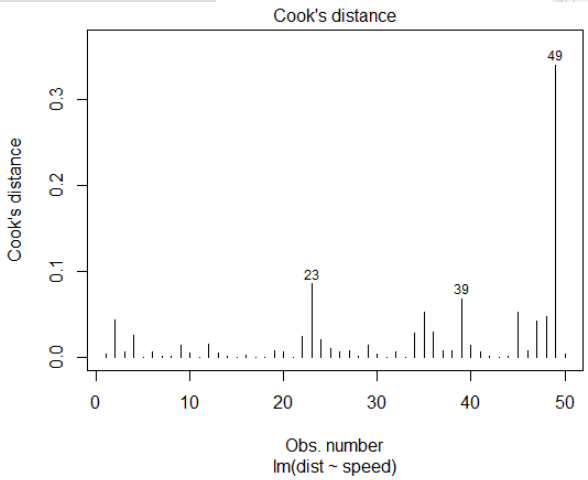
$$Y = aX + b + \mathcal{E}$$

где је $\mathcal{E}$ случајна величина независна и од Y и од X.

Случајна величина $\mathcal{E}$ често има $\mathcal{N}(0, \sigma^2)$ расподелу при чему σ^2 није познато. Из чињенице да је $E\mathcal{E} = 0$ следи закључак да су грешке мерења случајне, док је σ^2 у вези са карактеристикама мерног инструмента.

```
> #analiza reziduala
> #pretpostavka normalne raspodeljenosti reziduala
> shapiro.test(residuals(cars.lin.model)) #H0: reziduali su normalno raspodeljeni

        Shapiro-Wilk normality test

data:  residuals(cars.lin.model)
W = 0.9451, p-value = 0.02152
```

```
> #pretpostavka konstantnosti disperzije reziduala
> #H0: disperzija reziudala je ista za sve opservacije
> library(lmtest)
> bptest(cars.lin.model)

        studentized Breusch-Pagan test

data:  cars.lin.model
BP = 3.2149, df = 1, p-value = 0.07297

> #pretpostavka nezavisnosti reziduala
> #H0: autokorelacija reziduala je jednaka nuli
> dwtest(cars.lin.model, alternative="two.sided")

        Durbin-Watson test

data:  cars.lin.model
DW = 1.6762, p-value = 0.1904
alternative hypothesis: true autocorrelation is not 0
> #autlajeri
> sres <- rstandard(cars.lin.model) #racunaju se standardizovani reziduali, tj. reziduali podeljeni svojim standardnim devijacijama
> sres[which(abs(sres) > 2)] #ukazuje koje opservacije su autlajeri po y-vrednosti
   23    35    49
2.795 2.028 2.919
> leverage <- hatvalues(cars.lin.model) #odredjuju se "uticajne" vrednosti
> leverage[which(leverage > 4/50)] #leverages vece od 4/n, gde je n obim uzorka su sumnjive
      1       2      50
0.11486 0.11486 0.08727
> cooksD <- cooks.distance(cars.lin.model)
> F0.50 <- qf(0.5, df1=2, df=48)
> cooksD[which(cooksD > F0.50)]
named numeric(0)
> #nijedna opservacija nema ekstremno veliko Cook-ovo rastojanje
> plot(cars.lin.model, which=4) #vrednosti sa velikim Cook-ovim rastojanjem zasluzuju dodatno ispitivanje
```

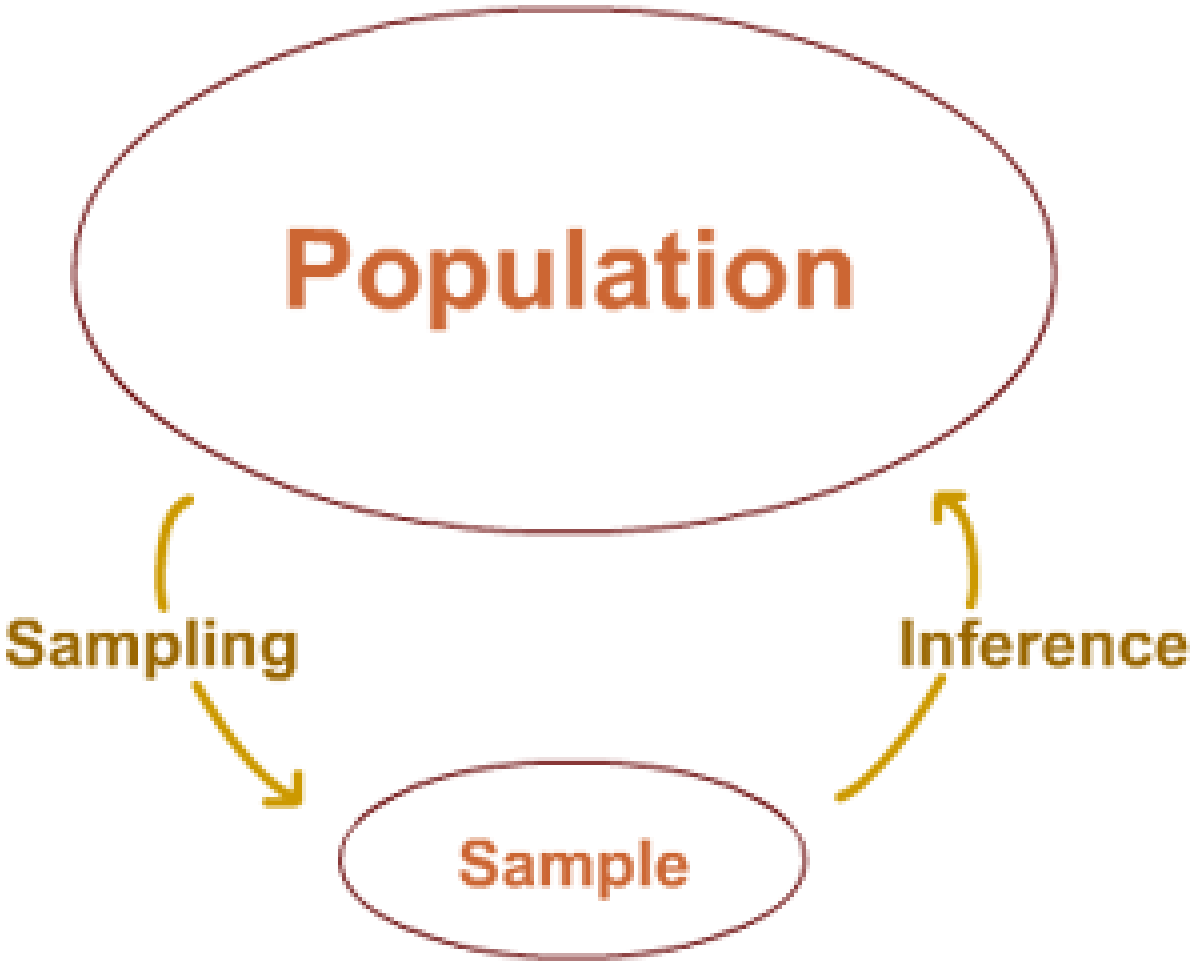


НАУЧНО ИСТРАЖИВАЊЕ

- Научно истраживање је систематско, планско и објективно испитивање неког проблема, према одређеним методолошким правилима, чија је сврха да се пружи поуздан и прецизан одговор на унапред постављено питање.
- Може се схватити као критички, контролисани и поновљиви процес стицања нових знања, неопходних (а понекад и довољних) за идентификовање, одређивање и решавање научних (теоријских и емпиријских) проблема.
- Емпиријско (искуствено) истраживање

- Свако научно истраживање има више међусобно логично повезаних фаза. Фазе су:
  - идентификовање и одређивање проблема
  - одређивање циља истраживања
  - постављање хипотезе
  - дефинисање кључних израза
  - извођење логичких последица из хипотезе
  - избор истраживачке стратегије и нацрта истраживања
  - развијање мерних и других средстава истраживања
  - одређивање основног скупа (популације) и одабирање узорка истраживања
  - спровођење истраживања и прикупљање значајних података
  - обрађивање и анализа података добијених истраживањем
  - тумачење резултата истраживања и извођење закључ(а)ка
  - писање извештаја о обављеном истраживању

- Трошкови (по питању уложеног времена, новца; очувања приватности и сл) прикупљања података на читавој популацији су обично прекомерни како за истраживаче тако и за испитиване објекте.
- Из поменутих разлога, у великој већини случајева, истраживањем не може бити обухваћена целокупна популација испитиваних објеката, него само део популације (узорак), па истраживач на основу налаза добијеног испитивањем узорка настоји да изведе закључак о целокупној популацији.
- Да би истраживач могао оправдано да уопштава налаз добијен испитивањем узорка, на популацију , неопходно је да буду испуњени неки, одређени услови.



ОСНОВНИ ПОЈМОВИ

- Коначна популација (*finite population*) је скуп/колекција, која садржи коначан број различитих елемената.
  - Елементи коначне популације су ентитети, који поседују одређене, заједничке карактеристике (оне су предмет интересовања истраживача). Елементи популације се другачије називају и јединице популације (*population units*).
- Обим/величина популације (*population size*) је број елемената коначне популације.
- Обично се означава са N, и увек је познат, коначан број.
  - Свакој јединици популације обима N придружује се (природан) број од 1 до N. Ти бројеви називају се ознаке јединица и они остају непромењени све до краја истраживања.

- Обележје је посматрана заједничка карактеристика елемената популације.
  - Вредности обележја y за јединице популације обима N означавају се са $Y_1, Y_2, \ldots, Y_N$. Овде Y_i означава вредност обележја y јединице означене са i.
- Параметар је реално-вредносна функција вредности обележја популације.

$$\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i \qquad S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \bar{Y})^2 \qquad R = \max_{1\le i\le N} Y_i - \min_{1\le i\le N} Y_i$$

  - Његова вредност (квантитативна) је често непозната, и о њој се закључује на основу информација добијених испитивањем узорка.
- Узорак (*sample*) је подскуп популације S.
  - Обично се означава са s.
  - Обим узорка је број елемената у узорку s, и означава се са $n(s)$.

- План узорковања (*sampling design*) је поступак којим се бирају елементи популације у узорак, уз одређивање адекватног обима узорка, а са циљем да се добије репрезентативан узорак и да се постигне максимална прецизност (тј. минимална дисперзија оцене посматраног обележја) по јединици трошкова.
- Статистика (*statistics*) је реално-вредносна функција, која зависи од $Y_1, Y_2, \dots, Y_N$ само преко s.
  - Када се статистика користи за оцењивање параметра она се назива оцена (*estimator*).

ПЛАНОВИ УЗОРКОВАЊА

- Вероватносно узорковање (*probability sampling*)
  - Свака оваква стратегија узорковања заснива се на теорији вероватноћа, при чему, у свакој етапи одабирања, вероватноћа ма ког скупа одабраних елемената популације мора бити позната. Дакле, узорковање се врши у складу са расподелом вероватноћа (*sampling design*) $\{P(s), s \in \Omega\}$, која је дефинисана на Ω (колекција свих могућих узорака).
  - Предности:
    - оцене параметара, базиране на статистикама, су непристрасне
    - постоји могућност одређивања грешке узорка
  - Стратегије/методи вероватносног узорковања:
    - прост случајан узорак
    - стратификован случајан узорак
    - систематски узорак
    - узорак скупина

- Невероватносно узорковање (*nonprobability sampling*)
  - Овакве стратегије узорковања не заснивају се на теорији вероватноћа. Њима се прибегава онда када је из разлога ограничених временских рокова, износа трошкова и етичких обзира тешко спровести случајно узорковање.
  - Ефикасно се примењују код експлоративних истраживања, чији циљ није прецизно оцењивање параметара на основу репрезентативног узорка.
  - Мане:
    - није могуће одређивање квалитета узорка, а самим тим ни тачности оцењивања
  - Стратегије невероватносног узорковања:
    - пригодни узорак
    - намерни узорак
    - квотни узорак
    - узорак “снежних грудви”

ЈОШ НЕКИ ОСНОВНИ ПОЈМОВИ

- Пристрасност (*bias*)
  - Нека је $P(.) = \{P(s), s \in \Omega\}$. Оцена $\hat{T}(.)$ је непристрасна за параметар θ у односу на $P(.)$, ако је

$$E_P\left[\hat{T}(s)\right] = \sum_{s\in\Omega} \hat{T}(s)P(s) = \theta$$

  - Разлика $E_P\left[\hat{T}(s)\right] - \theta$ назива се пристрасност $\hat{T}(.)$ при оцењивању θ у односу на $P(.)$.
- Средње квадратна грешка (*mean square error*)
  - Средње квадратна грешка оцене $\hat{T}(.)$ параметра θ у односу на $P(.)$ је

$$MSE\left(\hat{T}: P\right) = E_P\left[\hat{T}(s) - \theta\right]^2 = \sum_{s\in\Omega}\left[\hat{T}(s) - \theta\right]^2 P(s)$$

Треба приметити да се код непристрасне оцене, средње квадратна грешка своди на дисперзију.

Заправо, важи:

$$MSE(\hat{T}:P) = V_P(\hat{T}) + [B_P(\hat{T})]^2$$

где је $V_P(\hat{T})$ дисперзија, а $B_P(\hat{T})$ пристрасност статистике $\hat{T}(.)$

- Квалитет оцене вреднује се на бази њене пристрасности и средње квадратне грешке (треба бирати оцену која има мању пристрасност – ако је могуће чак да буде непристрасна, и мању средње кв. грешку).
- Ентропија (*entropy*)
  - Ентропија за дати $P(.)$ је

$$e = -\sum_{s\in\Omega} P(s)\ln P(s)$$

  - Како је ентропија мера информације у узорку, треба бирати $P(.)$ који има максималну ентропију.

- Индикатор укључења (*inclusion indicator*)
  - Нека је $\{s \ni i\}$ догађај да се узорак s садржи i-ту јединицу популације. Случајна величина

$$I_i(s) = \begin{cases} 1, & \text{ако је } s \ni i \\ 0, & \text{иначе} \end{cases}$$

    за $1 \leq i \leq N$, назива се индикатор укључења.
- Вероватноће укључења (*inclusion probabilities*)
  - Вероватноће укључења првог и другог реда за дати $P(.)$ су

$$\pi_i = \sum_{s \ni i} P(s) \quad \pi_{ij} = \sum_{s \ni i,j} P(s)$$

  - За дати $P(.)$ важи:

$$E_P[I_i(s)] = \pi_i, i = \overline{1,n}$$
$$E_P[I_i(s)I_j(s)] = \pi_{ij}, i,j = \overline{1,n}$$

У презентацији коришћена:

- књига “Статистичке методе у метеорологији и инжењерству” − Весна Јевремовић, Јован Малишић; Савезни хидрометереолошки завод, Београд 2002. год.

  и то, други део − математичка статистика, поглавље 19 − регресија и корелација (од стр. 217)