ТЕОРИЈА УЗОРАКА 1

5. 04. ‘13.

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:
  - 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 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 $\boldsymbol{N}$, 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 $Y_1, Y_2, \ldots, Y_N$ 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 $\{P(s), s \in \Omega\}$. Оцена $\hat{T}(\cdot)$ је непристрасна за параметар θ у односу на $P(\cdot)$, ако је

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

  - Разлика $E_P[\hat{T}(s)] - \theta$ назива се пристрасност $\hat{T}(\cdot)$ при оцењивању θ у односу на $P(\cdot)$.
  - Треба приметити да оцена која је непристрасна у односу на $P(\cdot)$ не мора бити непристрасна у односу на $Q(\cdot)$.
- **Средње квадратна грешка** (Mean Square Error)
  - Средње квадратна грешка оцене $\hat{T}(\cdot)$ параметра θ у односу на $P(\cdot)$ је

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

- 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 $\hat{T}(\cdot)$.

- Квалитет оцене вреднује се на бази њене пристрасности и средње кв. 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 $P(\cdot)$ је

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

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

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

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

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

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

  - За дати $P(\cdot)$ важи:

$$E_P[I_i(s)] = \pi_i \qquad E_P\left[I_i(s) I_j(s)\right] = \pi_{ij}$$

ПРОСТ СЛУЧАЈАН УЗОРАК (SIMPLE RANDOM SAMPLING)

- Ово је један од најједноставнијих и најстаријих метода бирања узорка обима n из популације која садржи N јединица.
- Нека је Ω колекција свих 2^N подскупова од S.

$$P(s) = \begin{cases} \binom{N}{n}^{-1}, & \text{ако је } n(s) = n \\ 0, & \text{иначе} \end{cases}$$

је simple random sampling design. У овом дизајну сваки од $\binom{N}{n}$ могућих скупова обима n се са подједнаком вероватноћом може одабрати као узорак.

- Поменути дизајн може се у пракси имплементирати следећим поступком извлачења јединица из популације:

  Одабрати n случајних бројева између 1 и N без понављања. Јединице популације које кореспондирају изабраним бројевима чине узорак.

- Код простог случајног узорка вероватноће укључења првог и другог реда једнаке су:

$$\pi_i = \frac{n}{N} \qquad \pi_{ij} = \frac{n(n-1)}{N(N-1)}$$

- Оцена тотала:

  Код простог случајног узорка $\hat{Y}_{srs} = \frac{N}{n}\sum_{i \in s} Y_i$ је

  непристрасна оцена тотала обележја популације, а њена дисперзија је

$$V\left[\hat{Y}_{srs}\right] = \frac{N^2(N-n)}{Nn} S_y^2$$

где је $S_y^2 = \frac{1}{N-1}\sum_{i=1}^{N}[Y_i - \bar{Y}]^2$.

- Непристрасна оцена $V[\hat{Y}_{srs}]$ је $v[\hat{Y}_{srs}] = \frac{N^2(N-n)}{Nn}s_y^2$ где је s_y^2 узорачки аналогон за S_y^2 .

- Ако је $\hat{\bar{Y}} = \frac{\hat{Y}_{srs}}{N}$ онда је, код простог случајног узорка, $\hat{\bar{Y}}$ непристрасна оцена средине обележја популације, а њена дисперзија је

$$\frac{N-n}{Nn}S_y^2$$

- **Фракција узорка** је количник $\frac{n}{N}$ и означава се са f.

СЛУЧАЈАН УЗОРАК СА ПОНАВЉАЊЕМ (RANDOM SAMPLING WITH REPLACEMENT)

- Разлика у односу на случајан узорак без понављања (прост случајан узорак) јесте у томе што се уместо дефинисања узорка као подскупа популације S (тада није дозвољено понављање јединица), узорак дефинише као низ чији су елементи јединице из S (тада узорак неће обавезно садржати различите јединице).
- $$P(s) = \begin{cases} \dfrac{n!}{N^n} \displaystyle\prod_{k \in S} \frac{1}{s_k!}, & \text{ако је } n(s) = n \\ 0, & \text{иначе} \end{cases}$$

  где је s_k број понављања k-те јединице у узорку.

ИНТЕРВАЛИ ПОВЕРЕЊА (CONFIDENCE INTERVALS)

- Када се одаберe узорак и на основу њега оцене параметри, пожељно је проценити тачност те оцене. То се обично постиже налажењем интервала поверења у оквиру којих се са довољно великом сигурношћу налазе популацијске вредности тотала или средине обележја, или, еквивалентно, одређивањем граница могућих грешака.

- Код простог случајног узорка:
  - $100(1-\alpha)\%$ интервал поверења за тотал обележја популације је

$$\hat{Y}_{srs} \pm t_{n-1;\frac{\alpha}{2}} \sqrt{\frac{N(N-n)}{n} s_y^2}$$

  - $100(1-\alpha)\%$ интервал поверења за средину обележја популације је

$$\hat{\bar{Y}} \pm t_{n-1;\frac{\alpha}{2}} \sqrt{\frac{N-n}{Nn} s_y^2}$$

    где је $t_{n-1;\frac{\alpha}{2}}$ вредност из таблица за Студентову расподелу са n-1 степени слободе, таква да је $P\left\{|t_n| \leq t_{n-1;\frac{\alpha}{2}}\right\} = 1-\alpha$

  - Ако је обим узорка већи од 30, вредност $t_{n-1;\frac{\alpha}{2}}$ се чита из таблица за стандардну нормалну расподелу.

ВЕЛИЧИНА УЗОРКА (SAMPLE SIZE)

- Код узорковања најчешће се поставља питање који обим узорка би требало одабрати. Одговор није увек једноставан.
- Код простог случајног узорка:
  - потребна величина узорка за оцену тотала обележја популације је

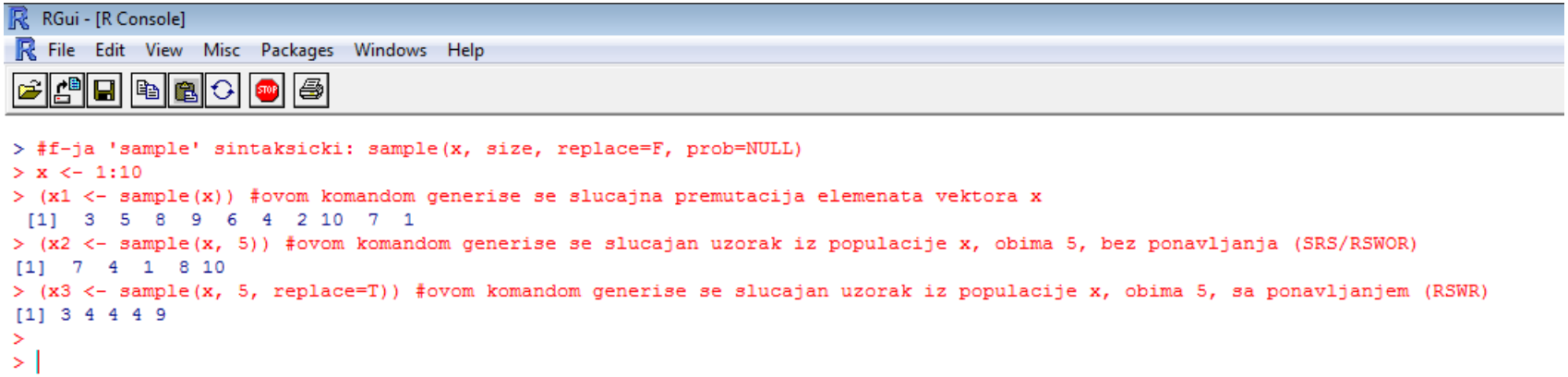
$$n = \frac{1}{\frac{1}{n_0} + \frac{1}{N}} \qquad n_0 = \frac{N^2 z^2 s_y^2}{d^2}$$

  - потребна величина узорка за оцену средине обележја популације је

$$n = \frac{1}{\frac{1}{n_0} + \frac{1}{N}} \qquad n_0 = \frac{z^2 s_y^2}{d^2}$$

  где је z вредност из таблица за стандардну нормалну расподелу, таква да је $P\{|Z| \leq z\} = 1 - \alpha$.

У R-у


RGui - [R Console]
File Edit View Misc Packages Windows Help
> #f-ja 'sample' sintaksicki: sample(x, size, replace=F, prob=NULL)
> x <- 1:10
> (x1 <- sample(x)) #ovom komandom generise se slucajna premutacija elemenata vektora x
[1] 3 5 8 9 6 4 2 10 7 1
> (x2 <- sample(x, 5)) #ovom komandom generise se slucajan uzorak iz populacije x, obima 5, bez ponavljanja (SRS/RSWOR)
[1] 7 4 1 8 10
> (x3 <- sample(x, 5, replace=T)) #ovom komandom generise se slucajan uzorak iz populacije x, obima 5, sa ponavljanjem (RSWR)
[1] 3 4 4 4 9
>