

Теорија узорака

24. март '20.

Како (нај)боље искористити податке?

- Приликом испитавања јединица у некој популацији, ради проучавања обележја Y од интереса, често је могуће узети у обзир различитост јединица на основу једног или више њихових додатних варијабилних својстава. Помоћне информације тог типа могу се користити или при пројектовању метода одабира узорка:
 - код узорковања са враћањем са неједнаким вероватноћама, где су вероватноће одабира пропорционалне „величини“ јединица узорковања
 - стратификација (раслојавање популације)

или **приликом закључивања** - у циљу „поправљања“ квалитета оцена:

- **количничко оцењивање**
- регресионо оцењивање

Количничко оцењивање SRSWOR

Нека је Y обележје од интереса. Нека је X **помоћно обележје** ('auxiliary variable', 'subsidiary variable'), које је у блиској вези - корелацији (суштински представља присуство приближне линеарне везе, али која пролази кроз координатни почетак) са „главним“ обележјем Y , и за које се претпоставља да се његова вредност може одредити на произвољној јединици у популацији. Мора бити познат и популацијски тотал τ_X обележја X .

За сваку јединицу у (простом) случајном узорку без понављања обима n из дате популације обима N регистроване су вредности и главног и помоћног обележја.

- Популацијски количник ('population ratio') B дефинисан је са:

$$B := \frac{\tau_Y}{\tau_X} = \frac{m_Y}{m_X},$$

а узорачки количник ('sample ratio') b са:

$$b = \frac{\sum_{k \in \mathcal{S}} y_k}{\sum_{k \in \mathcal{S}} x_k} = \frac{\bar{Y}}{\bar{X}}$$

Задатак је оцењивање непознатих параметара популације: B , m_Y или τ_Y .

Количничко оцењивање (наставак)

Кључни резултати у вези са количничким оцењивањем ('ratio estimation') неких непознатих популацијских вредности, када је **приступ заснован на методу** SRSWOR одабира узорка, дати су у следећој табели:

валидна само за „довољно“ велике узорке

тачкаста оцена $\hat{\theta}$ непознатог параметра θ	непристрасност	<u>апроксимација</u> дисперзије $D\hat{\theta}$	оцена дисперзије $D\hat{\theta}$	прилагођена оцена дисперзије $D\hat{\theta}$
$\hat{B} = b$	пристрасност: $\frac{-\text{cov}(b, \bar{X})}{m_X}$	$\frac{1}{m_X^2} \cdot \frac{\sigma_d^2}{n} \left(1 - \frac{n}{N}\right)$	$\frac{1}{m_X^2} \cdot \frac{\bar{S}_e^2}{n} \left(1 - \frac{n}{N}\right)$	$\frac{1}{\bar{X}^2} \cdot \frac{\bar{S}_e^2}{n} \left(1 - \frac{n}{N}\right)$
$\hat{m}_Y^r = b \cdot m_X$	пристрасност: $-\text{cov}(b, \bar{X})$	$\frac{\sigma_d^2}{n} \left(1 - \frac{n}{N}\right)$	$\frac{\bar{S}_e^2}{n} \left(1 - \frac{n}{N}\right)$	$\left(\frac{m_X}{\bar{X}}\right)^2 \cdot \frac{\bar{S}_e^2}{n} \left(1 - \frac{n}{N}\right)$

где су:

$$\sigma_d^2 := \frac{1}{N-1} \sum_{k=1}^N (y_k - Bx_k)^2$$

$$\bar{S}_e^2 := \frac{1}{n-1} \sum_{k \in S} (y_k - bx_k)^2$$

у пракси статистика $\bar{S}_e^2$ се може користити као тачкаста оцена σ_d^2 ;
њена пристрасност је реда величине $\frac{1}{n}$

Кључни **асимптотски резултат** у вези са интервалном оценом непознате популацијске средње вредности m_Y , добијеном применом количничког оцењивања, када је **приступ заснован на методу** SRSWOR:

апроксимативни $100 \cdot (1 - \alpha)\%$ двострани интервал поверења дат је са

$$\left[\hat{m}_Y^r - z_{1-\frac{\alpha}{2}} \sqrt{\frac{\bar{S}_e^2}{n} \left(1 - \frac{n}{N}\right)}, \hat{m}_Y^r + z_{1-\frac{\alpha}{2}} \sqrt{\frac{\bar{S}_e^2}{n} \left(1 - \frac{n}{N}\right)} \right]$$

када су n и $N - n$ „довољно велики“, уз задовољеност неких додатних услова

Резултати који се тичу непознатог популацијског тотала τ_Y лако се могу добити из резултата у вези са непознатом популацијском средњом вредношћу m_Y , коришћењем везе: $\tau_Y = Nm_Y$.

Количничко оцењивање

Количничко оцењивање даје добре резултате ако је веза између главног и помоћног обележја (приближно) линеарна, при чему одговарајућа права којом се описује ова зависност пролази кроз координатни почетак, и ако је распршеност вредности главног обележја око те праве пропорционална вредностима помоћног обележја. Када је **приступ заснован на моделу**, природно се намеће **линеарни регресиони модел** као модел популације: $x_1, x_2, \dots, x_N$ су познате константе (које су строго позитивне) а случајне величине $Y_1, Y_2, \dots, Y_N$ су **независне и дате са**:

$$Y_k = \beta x_k + \varepsilon_k,$$

где је $E\varepsilon_k (= E_M \varepsilon_k) = 0$ и $D\varepsilon_k (= D_M \varepsilon_k) = \sigma^2 x_k$, а β је фиксирана, али непозната вредност.

Нека је са $\mathcal{S}$ означен реализован (не)случајан узорак без понављања обима n . Ту имамо:

- најбоља линеарна оцена за β је:

$$\hat{\beta} = \frac{\sum_{k \in \mathcal{S}} Y_k}{\sum_{k \in \mathcal{S}} x_k}$$

- непристрасан (у односу на модел) линеарни предиктор случајне величине $\frac{1}{N} \sum_{k=1}^N Y_k$ са најмањом средње квадратном грешком дат је формулом:

$$\frac{1}{N} \left(\sum_{k \in \mathcal{S}} Y_k + \hat{\beta} \sum_{k \notin \mathcal{S}} x_k \right) = \hat{\beta} \cdot m_X$$

SRSWOR – оцењивање параметара потпопулације

Резултати добијени **применом количничког оцењивања** могу да се примене и код оцењивања непознатих потпопулацијских параметара ('estimation in (sub)domains'). Бира се (прост) случајан узорак без понављања $\mathcal{S}$ обима n из читаве популације обима N .

☞ Користе се исте ознаке као у презентацији за 3. час: N_l ; τ_l , m_l ; n_l ; $\mathcal{S}_l$

и уводе нове:

$$x_k = \begin{cases} 1, & \text{ако јединица означена са } k \text{ припада } l - \text{тој потпопулацији} \\ 0, & \text{иначе} \end{cases}$$
$$u_k = y_k x_k = \begin{cases} y_k, & \text{ако јединица означена са } k \text{ припада } l - \text{тој потпопулацији} \\ 0, & \text{иначе} \end{cases}$$

Тада је: $\tau_X = N_l$, $m_X = \frac{N_l}{N}$, $\tau_U := \sum_{k=1}^N u_k = \tau_l$, $m_U := \frac{\tau_U}{N_l} \left(= \frac{\tau_U}{\tau_X} =: B \right) = m_l$, $\bar{X} = \frac{\sum_{k \in \mathcal{S}} x_k}{n} = \frac{n_l}{n}$ и

$$\hat{m}_l = \hat{B} = \frac{\sum_{k \in \mathcal{S}} u_k}{\sum_{k \in \mathcal{S}} x_k}$$

Прилагођена оцена стандардне грешке оцене $\hat{m}_l$ дата је формулом (слајд 4):

$$\sqrt{\frac{n}{n-1} \cdot \frac{n_l-1}{n_l^2} \left(1 - \frac{n}{N}\right) \bar{s}_{le}^2}, \quad \text{где је } \bar{s}_{le}^2 := \frac{1}{n_l-1} \sum_{k \in \mathcal{S}_l} (y_k - \hat{B})^2$$

Количничко оцењивање – узорак са неједнаким вероватноћама

Уопштена количничка оцена непознатог популацијског тотала τ_Y , када је приступ заснован на методу случајног узорка са понављањем, односно без понављања, и неједнаким вероватноћама укључења, дата је са:

$$\hat{t}_G = \frac{\hat{t}_Y}{\hat{t}_X} \cdot \tau_X,$$

где су $\hat{t}_Y$ и $\hat{t}_X$ Horvitz-Thompson-ове оцене, редом, τ_Y и τ_X , тј.

$$\hat{t}_Y = \sum_{k \in \mathcal{S}'} \frac{y_k}{\pi_k} \quad \text{и} \quad \hat{t}_X = \sum_{k \in \mathcal{S}'} \frac{x_k}{\pi_k},$$

$\mathcal{S}'$ је редуковани узорак, који садржи само различите јединице из оригиналног узорка $\mathcal{S}$, а π_k су (познате) вероватноће укључења првог реда.

- Пример - количничко оцењивање - SRSWOR

Влада САД сваких пет година спроводи Попис пољопривреде, прикупљајући податке о свим фармама (фарма се дефинише као било које имање на коме је (односно, би, под уобичајеним околностима, било), током пописне године, произведено и продато пољопривредних производа у вредности од бар \$1000) на територији 50 држава.

Попис пољопривреде садржи податке о броју пољопривредних газдинстава, укупној површини фарми, приносу различитих врста усева и разним другим пољопривредним показатељима за свих $N = 3078$ округа, односно области еквивалентних окрузима, у САД.

оквир - Попис пољопривреде из 1992. г. * ;
дати су подаци за 1982, 1987 и 1992. г. о површини фарми, броју фарми, броју „малих“ фарми - површине мање од 9 јутара, броју „великих“ фарми - површине веће од 1000 јутара, за комплетну популацију (цензус)

$n = 300$, узорак се бира методом SRSWOR коришћењем програмског генератора **

† јединице узорковања су окрузи

главно обележје Y - укупна површина фарми 1992. г.

помоћно обележје X - укупна површина фарми 1987. г.

* подаци се налазе у бази агрор у оквиру  пакета SDaA

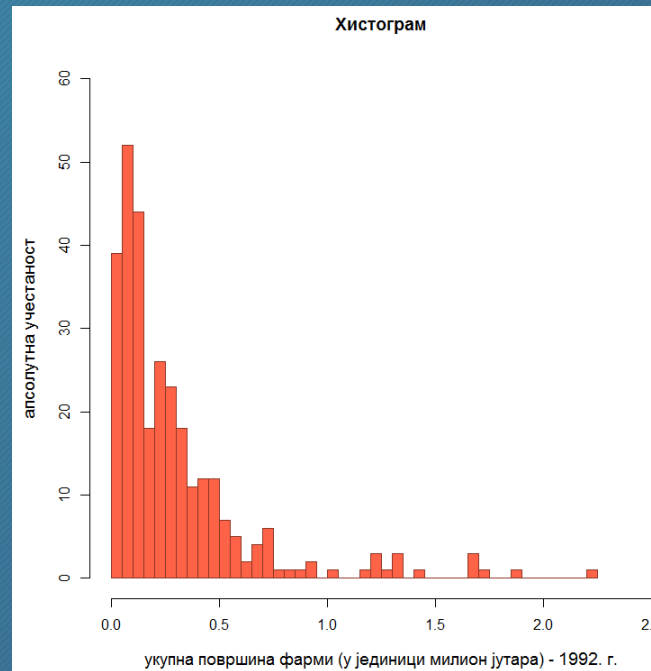
** узорак се налази у бази agsrs у истом пакету

- Пример - количничко оцењивање - SRSWOR - наставак

Узорачка средина главног обележја је: $\bar{Y} = 297897$, па је $N\bar{Y} = 916927110$.

Оцењене стандардне грешке статистике $\bar{Y}$, односно $N\bar{Y}$, су, редом: 18898.43, односно 58169381.

На следећем графику дат је хистограм вредности главног обележја из узорка:



На основу познавања цензуса за 1987. г. (👉 ово се не може одредити директно из базе, јер у њој недостају / нису доступни подаци о вредностима помоћног обележја за 23 округа) лако се добија:

$\tau_X = 964470625$ јутара, односно $m_X = 313343.3$ јутара (по округу)

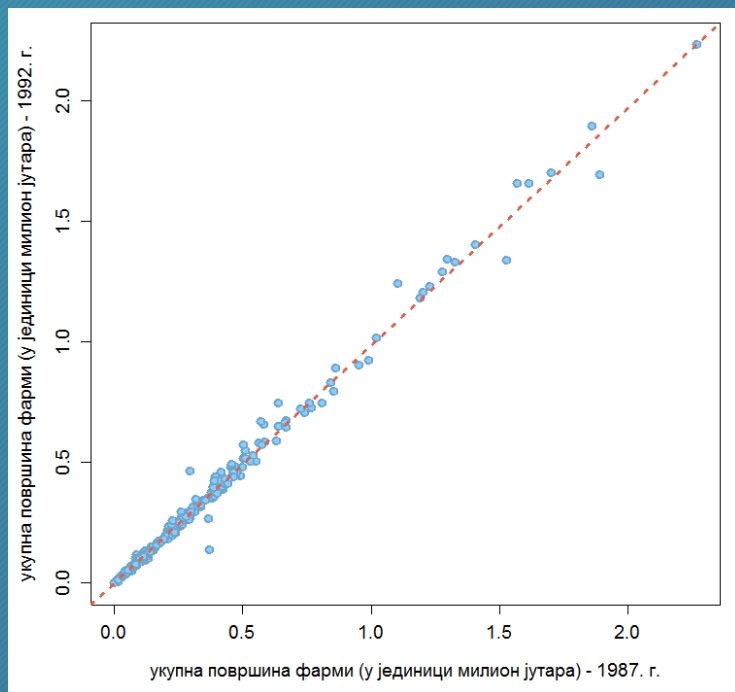
Узорачка средина помоћног обележја је: $\bar{X} = 301953.7$

- Пример - количничко оцењивање - SRSWOR - наставак

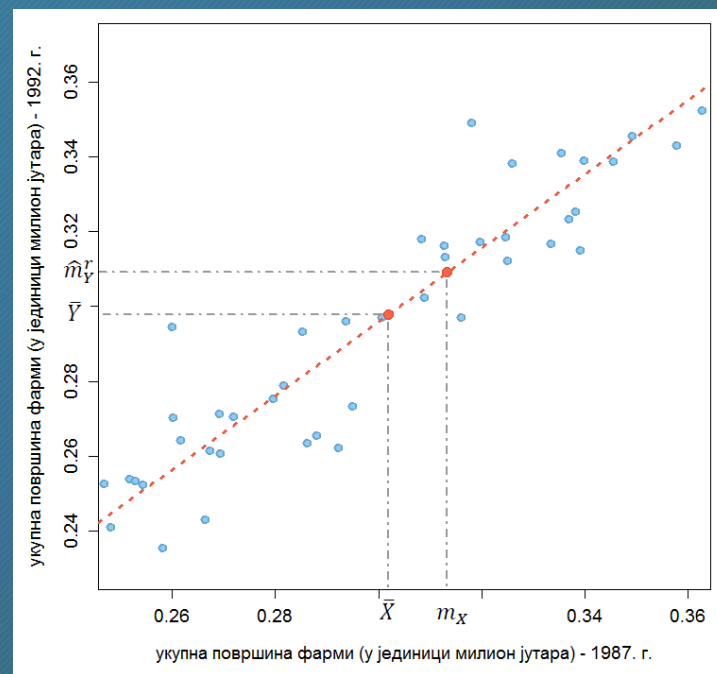
Оцењена вредност

- количника је: $\hat{B} \left(= \frac{\bar{Y}}{\bar{X}} \right) = 0.9865652$
- популацијског тотала је: $\hat{\tau}_Y^r (= \hat{B}\tau_X) = 951513191$

Оцењена вредност (прилагођена оцена) стандардне грешке оцено $\hat{\tau}_Y^r$ је: 5546162



На графику је приказан дијаграм распршености података из узорка, тј. тачака (x_k, y_k) , $k \in S$. Такође је уцртана права која пролази кроз координатни почетак и има коефицијент правца једнак $\hat{B}$.



Увећан делић графика са леве стране.

За одабрани узорак:

$$m_X > \bar{X},$$

па је

$$\hat{m}_Y^r > \bar{Y}$$

Како су вредности главног и помоћног обележја позитивно корелиране, потцењена оцена $\bar{X}$ тачне популацијске средине обележја X указује на то да ће $\bar{Y}$ такође потценити популацијску средњу вредност обележја Y . Количничко оцењивање даје прецизнију оцено m_Y , проширујући $\bar{Y}$ фактором $m_X/\bar{X}$.