

Теорија узорака

5. мај '20.

- Пример - групни узорак

Једног полазника кошаркашког кампа за дечаке узраста 15-17 г. интересовала је просечна висина дечака са којима борави у кампу. Укупан број полазника кампа био је 100 и они су били смештени у петокреветним собама.

Дечак се определио да, прво, на случајан начин и без враћања из качкета извуче четири цедуљице на којима су записани бројеви соба и пита сваког полазника из тих соба колика је његова висина. Затим је замолио тренера да му да списак са именима свих полазника, поређаним по азбучном реду, како би на основу њега (као оквира) одабрао (прост) случајан узорак без понављања, односно периодичан узорак, обима 20 дечака.

Дечак је добио следеће резултате у вези са оцењивањем популацијске средње вредности за секундарну јединицу (тј. за појединачног полазника) обележја Y - висина (у m):

- групни узорак

оцењена вредност популацијске средине: $\hat{m}_Y^{clu} = 1.721 m$

оцењена вредност стандардне грешке оцено $\hat{m}_Y^{clu}$: $0.0124043 m$

- (прост) случајан узорак без понављања

оцењена вредност популацијске средине: $\hat{m}_Y = \bar{Y} = 1.772 m$

оцењена вредност стандардне грешке оцено $\hat{m}_Y$: $0.01383512 m$

- периодичан узорак

оцењена вредност популацијске средине: $\hat{m}_Y^{sys} = 1.7615 m$

оцењена вредност стандардне грешке оцено $\hat{m}_Y^{sys}$: $0.01410151 m$

$$m_Y = 1.7423 m$$

- Пример - групни узорак

Швајцарски Федерални завод за статистику (Swiss Federal Statistical Office) спроводи испитивања националног цензуса са циљем пружања релевантних закључака о тренутном стању и развоју становништва, економије, друштва, образовања, истраживања, територије и животне средине у Швајцарској. Федерални попис становништва даје вредне информације из ових области.

оквир - база података *; у њој се налазе регистровани подаци из 2003. г. за 22 обележја за сваку од 2896 швајцарских општина

обележје Y - број становника са 65+ година живота

$n = 300$ општина, случајан узорак са враћањем бира се са вероватноћама пропорционалним бројевима становника општина

Оцењена вредност укупног броја становника са 65+ година живота је:

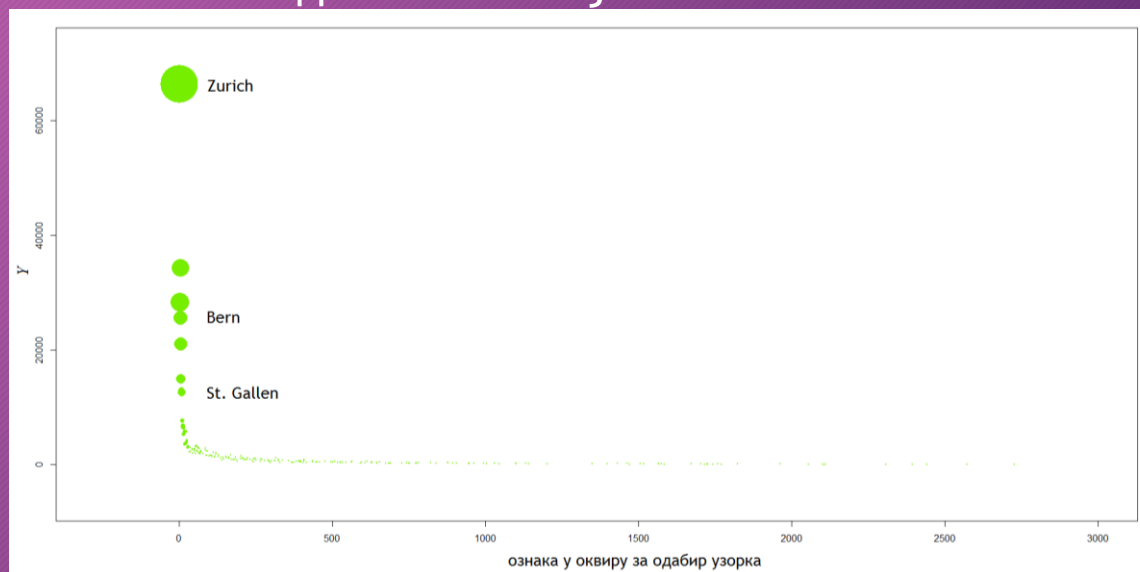
- $\hat{t}_{\psi} = 1\,126\,401$ Н-Н оцена
- $\hat{t}_{\pi} = 1\,086\,226$ Н-Т оцена

Тачна вредност је: $\tau_Y = 1\,119\,006$

Оцењена стандардна грешка једнака је:

- 15342.09 за Н-Н оцену
- 31853.87 за Н-Т оцену

Редуковани узорак
има обим: 224 општине

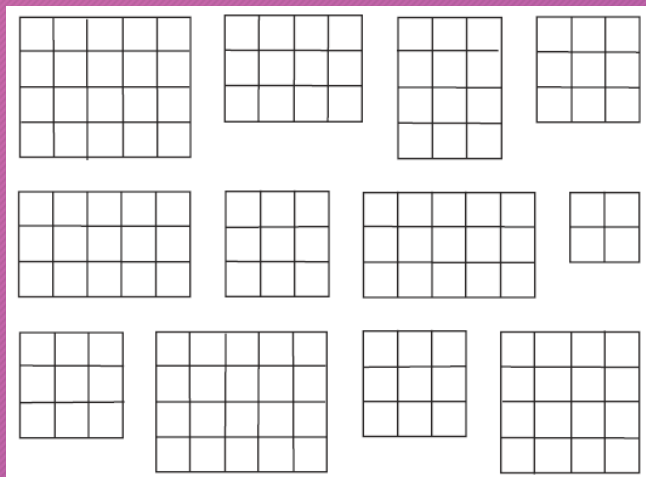


Вишеетапни групни узорак

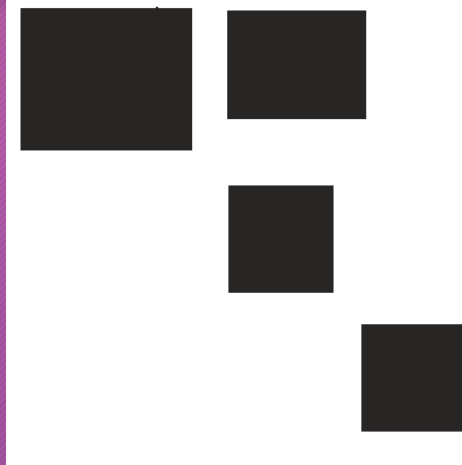
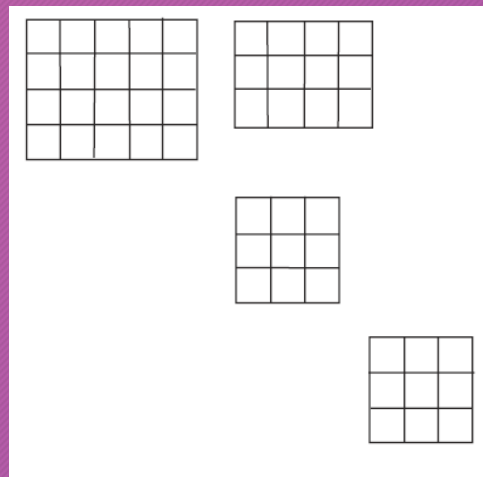
- Једноетапни групни узорак је по структури веома једноставан, али када су обими група велики може бити **непрактичан**. Такође, у многим ситуацијама, ентитети који сачињавају једну групу могу бити толико слични да испитивање СВИХ њих представља чисто **разбацивање ресурса**, односно регистровање вредности обележја на секундарним јединицама може бити **скупо** у поређењу са трошком узорковања групе.
- Код **вишеетапног групног узорка** ('multistage cluster sampling') подела комплетне популације извршена је на тзв:
 - ▼ примарне јединице
 - ▼ секундарне јединице
 - ▼ терцијарне јединице
 - ▼ итд.
- Јединице посматрања (ентитети) су овде елементи „најситније“ поделе у горњој хијерархији, док су јединице узорковања, по етапама, елементи одговарајуће поделе (по вертикали, одозго надоле).

Двоетапни једноетапни групни узорак

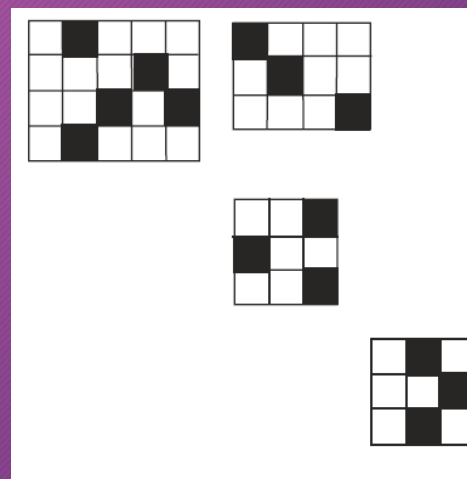
Популација подељена на
 N примарних јединица



Одабир (простог) случајног
узорка без понављања
примарних јединица



Испитивање свих ентитета
из узорка група –
једноетапни групни узорак



Одабир (простог) случајног
узорка без понављања
ентитета из сваке примарне
јединице из узорка група –
двоетапни групни узорак

Ознаке

- N – број група; M_l – број секундарних јединица (ентитета) у l -тој групи, $l = \overline{1, N}$

$$M = \sum_{l=1}^N M_l \text{ – обим популације}$$

- n – обим узорка $\mathcal{S}$ група, тј. примарних јединица
- n_l – обим узорка $\mathcal{S}_l$ ентитета, тј. секундарних јединица из l -те групе, $l = \overline{1, N}$
- y_{lk} – вредност обележја Y ентитета означеног са k , која потиче из l -те групе
- вредности у вези са појединачном групом:

унапред
познато

- популацијске вредности:

τ_l		m_l
τ_Y	μ_Y	m_Y
σ_τ^2		σ_l^2

SRSWOR метод одабира у свакој етапи

- Непристрасна оцена

Приступ приликом закључивања заснован је на методу одабира узорка. Кључни резултати дати су у следећој табели:

тачкаста оцена $\hat{\theta}$ непознатог параметра θ	дисперзија $D\hat{\theta}$	оцена $\widehat{D\hat{\theta}}$ дисперзије $D\hat{\theta}$
$\hat{t}_l = \frac{M_l}{n_l} \sum_{k \in \mathcal{S}_l} y_{lk}$ $= M_l \cdot \bar{Y}_l$		
$\pi_{lk} = \frac{n}{N} \cdot \frac{n_l}{M_l} \leftarrow \hat{t}_Y^u = \frac{N}{n} \sum_{l \in \mathcal{S}} \hat{t}_l = \sum_{l \in \mathcal{S}} \sum_{k \in \mathcal{S}_l} \frac{N}{n} \cdot \frac{M_l}{n_l} \cdot y_{lk}$	$\frac{N^2}{n} \left(1 - \frac{n}{N}\right) \sigma_\tau^2 + \frac{N}{n} \sum_{l=1}^N \frac{M_l^2}{n_l} \left(1 - \frac{n_l}{M_l}\right) \sigma_l^2$	$\frac{N^2}{n} \left(1 - \frac{n}{N}\right) \bar{S}_\tau^2 + \frac{N}{n} \sum_{l \in \mathcal{S}} \frac{M_l^2}{n_l} \left(1 - \frac{n_l}{M_l}\right) \bar{S}_l^2$
$\hat{m}_Y^u = \frac{\hat{t}_Y^u}{M}$	$\frac{D\hat{t}_Y^u}{M^2}$	$\frac{\widehat{D\hat{t}_Y^u}}{M^2}$

где су $\bar{Y}_\tau$ и $\bar{S}_\tau^2$, $\bar{S}_l^2$, редом, дати формулама:

$$\bar{Y}_\tau = \frac{\hat{t}_Y^u}{N} \text{ и } \bar{S}_\tau^2 = \frac{1}{n-1} \sum_{l \in \mathcal{S}} (\hat{t}_l - \bar{Y}_\tau)^2, \quad \bar{S}_l^2 = \frac{1}{n_l-1} \sum_{k \in \mathcal{S}_l} (y_{lk} - \bar{Y}_l)^2$$

SRSWOR метод одабира у свакој етапи (наставак)

- Количничка оцена

Приступ приликом закључивања заснован је на методу одабира узорка. Кључни резултати дати су у следећој табели:

тачкаста оцена $\hat{\theta}$ непознатог параметра θ	непристрасност	апроксимација дисперзије $D\hat{\theta}$	оцена $\widehat{D\theta}$ дисперзије $D\hat{\theta}$
$\hat{\tau}_Y^r = \hat{b} \cdot M$	×	$\frac{N^2}{n} \left(1 - \frac{n}{N}\right) \cdot \frac{1}{N-1} \sum_{l=1}^N (\tau_l - m_Y \cdot M_l)^2$ $+ \frac{N}{n} \sum_{l=1}^N \frac{M_l^2}{n_l} \left(1 - \frac{n_l}{M_l}\right) \sigma_l^2$	$\frac{N^2}{n} \left(1 - \frac{n}{N}\right) \cdot \frac{1}{n-1} \sum_{l \in \mathcal{S}} (\hat{\tau}_l - \hat{b} \cdot M_l)^2$ $+ \frac{N}{n} \sum_{l \in \mathcal{S}} \frac{M_l^2}{n_l} \left(1 - \frac{n_l}{M_l}\right) \bar{S}_l^2$

где је:

$$\hat{b} := \frac{\sum_{l \in \mathcal{S}} \hat{\tau}_l}{\sum_{l \in \mathcal{S}} M_l}$$

PPS метод одабира група

Hansen-Hurwitz-ова оцена

Претпоставља се да је одабран узорак група обима n са понављањем, при чему је за фиксирано l позната вероватноћа ψ_l одабира l -те групе. Претпоставимо да је она дата са:

$$\psi_l = \frac{M_l}{M}.$$

Затим се из сваке примарне јединице, која је одабрана у узорак група, (независно од осталих) бира вероватносни узорак ентитета, унапред задатог обима.

👉 У тој, другој, етапи обично се примењује метод простог случајног узорка без понављања или систематско узорковање секундарних јединица.

Претпоставља се да су у другој етапи ентитети одабрани применом SRSWOR метода.

Уведу се случајне величине Q_l , $l = \overline{1, N}$, где Q_l представља број појављивања групе означене са l у узорку група. Ако је l -та група одабрана у узорак више од једанпут, онда постоји **тачно Q_l независних оцена** $\hat{t}_{l1}, \hat{t}_{l2}, \dots, \hat{t}_{lQ_l}$ тотала обележја Y за l -ту групу.

Приступ приликом закључивања заснован је на методу одабира узорка. Кључни резултати дати су у следећој табели:

тачкаста оцена $\hat{\theta}$ непознатог параметра θ	непристрасност	оцена $\widehat{D\hat{\theta}}$ дисперзије $D\hat{\theta}$
$\hat{t}_Y^\psi = \frac{M}{n} \sum_{l=1}^N \sum_{j=1}^{Q_l} \frac{\hat{t}_{lj}}{M_l}$	✓	$\frac{M^2}{n(n-1)} \sum_{l=1}^N \sum_{j=1}^{Q_l} \left(\frac{\hat{t}_{lj}}{M_l} - \frac{\hat{t}_Y^\psi}{M} \right)^2$

Horvitz-Thompson-ова оцена

Нека је $\mathcal{S}'$ редуковани узорак група и нека су познате вероватноће укључења првог реда π_l за сваку групу.

Приступ приликом закључивања заснован је на методу одабира узорка.

Кључни резултат је:

тачкаста оцена $\hat{\theta}$
непознатог параметра θ

$$\hat{t}_Y^\pi = \sum_{l \in \mathcal{S}'} \frac{\hat{t}_l}{\pi_l}$$

Планирање двоетапног групног узорковања

Додатно питање:

- Колики би требало да буде обим узорка секундарних јединица из појединачне примарне јединице узорковане у првој етапи?

Одређивање обима (под)узорка

Главни циљ планирања истраживања јесте: добити што је могуће више информација уз најмањи (не)материјални трошак.

Ради једноставности, у наставку претпоставимо да је свака од N група сачињена од подједнаког броја ентитета, тј. $M_l = \bar{M} = \frac{M}{N}$, $l = \overline{1, N}$.

Посматра се непристрасна оцена $\hat{t}_Y^u$ непознатог популацијског тотала добијена на основу двоетапног узорка, одабраног применом SRSWOR метода у обе етапе. Обим узорка примарних јединица је n , а обим подузорка секундарних јединица из сваке од њих је $\bar{n}$.

Једноставна функција којом се може описати (укупан) трошак истраживања је:

$$C = C_0 + C_1 \cdot n + C_2 \cdot n\bar{n},$$

где је C_0 општи трошак, C_1 трошак по одабраној групи (не укључујући трошак испитивања ентитета унутар ње) и C_2 трошак по одабраном ентитету.

Вредност $\bar{n}$ обима подузорка (ентитета), која минимизује дисперзију

$$D\hat{t}_Y^u = \frac{M^2}{n\bar{M}} \left(1 - \frac{n}{N}\right) \text{MSB} + \frac{M^2}{n\bar{n}} \left(1 - \frac{\bar{n}}{\bar{M}}\right) \text{MSW}$$

за познато и фиксирано C , дата је формулом:

$$\bar{n}_{\text{opt}} = \sqrt{\frac{C_1 \cdot \bar{M} \cdot \text{MSW}}{C_2 (\text{MSB} - \text{MSW})}}.$$

Вредност обима n узорка (група) тада је:

$$n = \frac{C - C_0}{C_1 + C_2 \cdot \bar{n}_{\text{opt}}}.$$

Одређивање обима (под)узорка (наставак)

Сматра се да су дефинисане примарне јединице и одређене њихове величине, као и да је установљен разломак подузорковања $\frac{n_l}{M_l}$ ентитета по групама. Задатак је одређивање обима n узорка група за оцењивање непознате популацијске средње вредности m_Y , тако да се постигне апсолутна грешка не већа од Δ са поверењем $1 - \alpha$.

Ради једноставности, у наставку претпоставимо да је свака од N група сачињена од подједнаког броја ентитета, тј. $M_l = \bar{M} = \frac{M}{N}$, $l = \overline{1, N}$.

Користи се:

$$D\hat{m}_Y^u \leq \frac{1}{n} \left(\frac{\text{MSB}}{\bar{M}} + \frac{\text{MSW}}{\bar{n}} \left(1 - \frac{\bar{n}}{\bar{M}} \right) \right) = \frac{v}{n}$$

па је апроксимативни $100 \cdot (1 - \alpha)\%$ интервал поверења за m_Y дат са:

$$\left[\hat{m}_Y^u - z_{1-\frac{\alpha}{2}} \sqrt{\frac{v}{n}}, \hat{m}_Y^u + z_{1-\frac{\alpha}{2}} \sqrt{\frac{v}{n}} \right].$$

Тражена („груба“) процена вредности обима узорка n може се дати формулом:

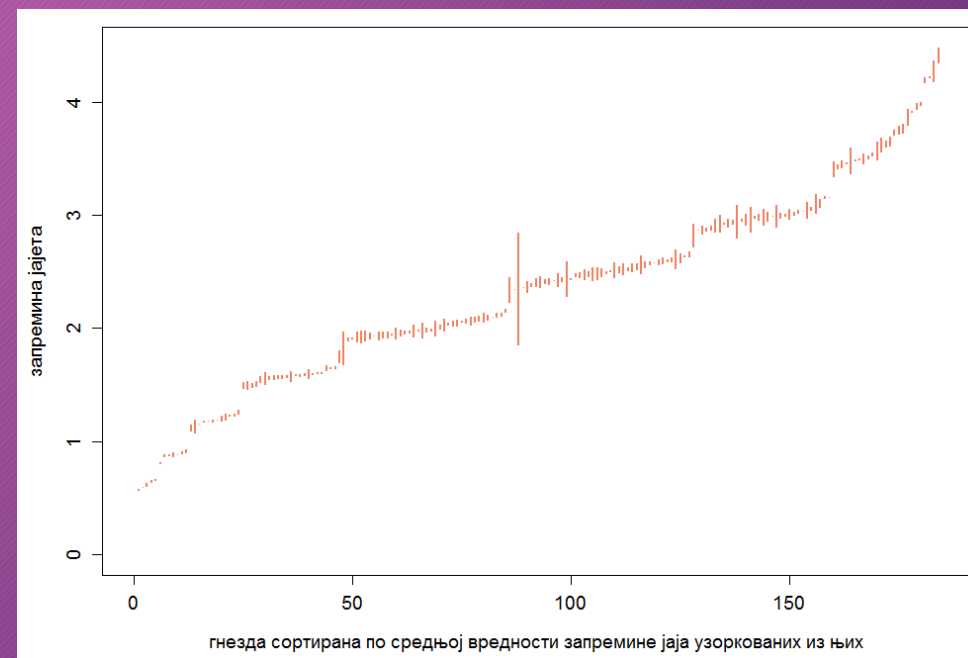
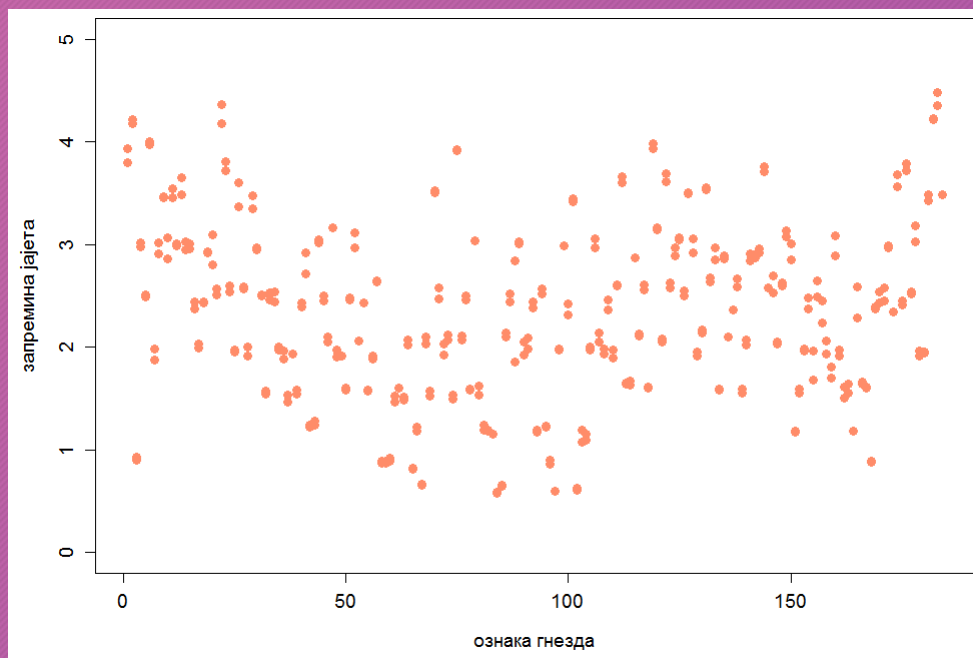
$$n = \frac{v \left(z_{1-\frac{\alpha}{2}} \right)^2}{\Delta^2}.$$

- Пример - двоетапни групни узорак



Дати су подаци из Arnold-овог истраживања (1991. г), чији је предмет било проучавање величине и **запремине јајета** америчке лиске (врста птице из породице барских кока, на територији насеља Минедоса (Minnedosa) у канадској провинцији Манитоба (Manitoba).

Подаци за подзорке обима 2 јајета из узорка обима 184 гнезда са бар два јајета доступни су у бази *.



* coots у оквиру  пакета SDaA

- Пример - двоетапни групни узорак – наставак

За посматрано обележје Y - запремина јајета америчке лиске, на основу двоетапног групног узорка, добијени су следећи закључци:

оцењена вредност (количничка оцена) популацијске средине m_Y је: $\hat{m}_Y^r = \hat{b} = 2.490417$
оцењена вредност стандардне грешке оцено $\hat{m}_Y^r$ је: 0.06097515

☞ Оцењена вредност стандардне грешке оцено $\hat{m}_Y^r$ добијена је тако што је, у формули са слајда 8, потпуно одбачен други сабирак као занемарљиво мали у односу на први сабирак,

а у првом сабирку стављено $1 - \frac{n}{N} \approx 1$

(сматрајући да је број гнезда N у популацији велики (иако непознат, свакако значајно већи од $n = 184$))

и просечна величина групе $\bar{M} := \frac{M}{N}$, такође, оцењена на основу узорка $\mathcal{S}$ група са $\frac{\sum_{l \in \mathcal{S}} M_l}{n}$.