

# Теорија узорака

3. март '20.

# Узорак са неједнаким вероватноћама

- Код случајног узорка са неједнаким вероватноћама ('unequal probability sampling') често:

јединица посматрања  $\neq$  јединица узорковања.

Нека је дата популација са  $N$  јединица, које су у оквиру за одабир узорка означене бројевима из скупа  $\Omega = \{1, 2, \dots, N\}$  и нека је  $Y$  обележје од интереса, при чему  $y_k$  представља вредност тог обележја јединице означене са  $k$ . Узорак обима  $n$  бира се сукцесивно - извлачењем једне по једне јединице са враћањем (тј. са понављањем) или без враћања (тј. без понављања).

- Претпоставља се да су, за свако  $k = \overline{1, N}$ , **познате**: вероватноћа укључења првог реда  $\pi_k$ , односно  $\psi_k$  – **вероватноћа одабира** јединице означене са  $k$  у првом извлачењу.

# СА ПОНАВЉАЊЕМ

Случајан узорак  $\mathcal{R}$  у наставку сматрамо неуређеним скупом кардиналности  $n$ , са евентуалним понављањима ознака јединица. За фиксирано  $k$  вероватноћа  $\psi_k$  представља вероватноћу да у произвољном извлачењу (другом, ...,  $n$ -том) јединица означена са  $k$  буде одабрана у узорак. За вероватноћу укључења првог реда важи:

$$\pi_k = 1 - (1 - \psi_k)^n$$

Кључни резултати у вези непознатим популацијским тоталом  $\tau_Y$ , када је приступ заснован на методу случајног узорка са понављањем и неједнаким вероватноћама одабира (тј. укључења првог реда) дати су у следећој табели:

$$\pi_{kl} = \pi_k + \pi_l - 1 + (1 - \psi_k - \psi_l)^n$$

|                                 |                                                                                                |
|---------------------------------|------------------------------------------------------------------------------------------------|
| тачкаста оцена $\hat{t}_\psi$   | $\frac{1}{n} \sum_{k \in \mathcal{R}} \frac{y_k}{\psi_k}$                                      |
| $E \hat{t}_\psi$                | $\tau_Y$                                                                                       |
| тачкаста оцена $D \hat{t}_\psi$ | $\frac{1}{n(n-1)} \sum_{k \in \mathcal{R}} \left( \frac{y_k}{\psi_k} - \hat{t}_\psi \right)^2$ |

Hansen-Hurwitz-ова оцена

# СА / БЕЗ ПОНАВЉАЊА

Нека су познате вероватноће укључења првог реда  $\pi_k$  и вероватноће укључења другог реда  $\pi_{kl}$ ; за све њих претпоставља се да су строго позитивне.

Кључни резултати у вези са непознатим популацијским тоталом  $\tau_Y$ , када је приступ заснован на методу случајног узорка са понављањем, односно без понављања, и неједнаким вероватноћама укључења дати су у следећој табели:

|                                |                                                                                                                                                                                                                                     |
|--------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| тачкаста оцена $\hat{t}_\pi$   | $\sum_{k \in \mathcal{S}'} \frac{y_k}{\pi_k}$                                                                                                                                                                                       |
| $E \hat{t}_\pi$                | $\tau_Y$                                                                                                                                                                                                                            |
| тачкаста оцена $D \hat{t}_\pi$ | $\sum_{k \in \mathcal{S}'} \left( \frac{1}{\pi_k^2} - \frac{1}{\pi_k} \right) y_k^2 + \sum_{k \in \mathcal{S}'} \sum_{\substack{l \in \mathcal{S}' \\ l \neq k}} \left( \frac{1}{\pi_k \pi_l} - \frac{1}{\pi_{kl}} \right) y_k y_l$ |

Horvitz-Thompson-ова оцена

→ може се десити да се добије негативна вредност као оцена дисперзије

где је  $\mathcal{S}'$  редуковани узорак, тј. неуређен скуп кардиналности  $n_D$  у коме су изостављена евентуална понављања јединица из оригиналног узорка,  $n_D$  представља ефективан обим узорка

# БЕЗ ПОНАВЉАЊА

Нека је ефективан обим узорка  $n_D = n$  фиксирана (неслучајна) вредност (заправо, имамо случајан узорак  $\mathcal{S}$  без понављања обима  $n$ ).

Кључни резултати у вези непознатим популацијским тоталом  $\tau_Y$  дати су у следећој табели:

|                                |                                                                                                                                                                                       |
|--------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| тачкаста оцена $\hat{t}_\pi$   | $\sum_{k \in \mathcal{S}} \frac{y_k}{\pi_k}$                                                                                                                                          |
| $E \hat{t}_\pi$                | $\tau_Y$                                                                                                                                                                              |
| тачкаста оцена $D \hat{t}_\pi$ | $\frac{1}{2} \sum_{k \in \mathcal{S}} \sum_{\substack{l \in \mathcal{S} \\ l \neq k}} \frac{\pi_k \pi_l - \pi_{kl}}{\pi_{kl}} \left( \frac{y_k}{\pi_k} - \frac{y_l}{\pi_l} \right)^2$ |

Sen-Yates-Grundy-јева оцена

Резултати који се тичу непознате популацијске средње вредности  $m_Y$  лако се могу добити из резултата у вези са непознатим популацијским тоталом  $\tau_Y$ , коришћењем везе:  $\tau_Y = Nm_Y$ .

Међутим, Horvitz-Thompson-ова оцена, иако непристрасна може имати велику дисперзију када се примењује на оне популације код којих није „добра“ веза између вредности обележја од интереса и вероватноћа укључења (тј.  $\pi_k$  и  $y_k$  нису приближно пропорционални). Тада се уводи уопштена оцена непознате популацијске средње вредности  $m_Y$  дата са:

$$\hat{m}_Y = \frac{\sum_{k \in \mathcal{S}'} \frac{y_k}{\pi_k}}{\sum_{k \in \mathcal{S}'} \frac{1}{\pi_k}}$$

Најѐк-ова оцена

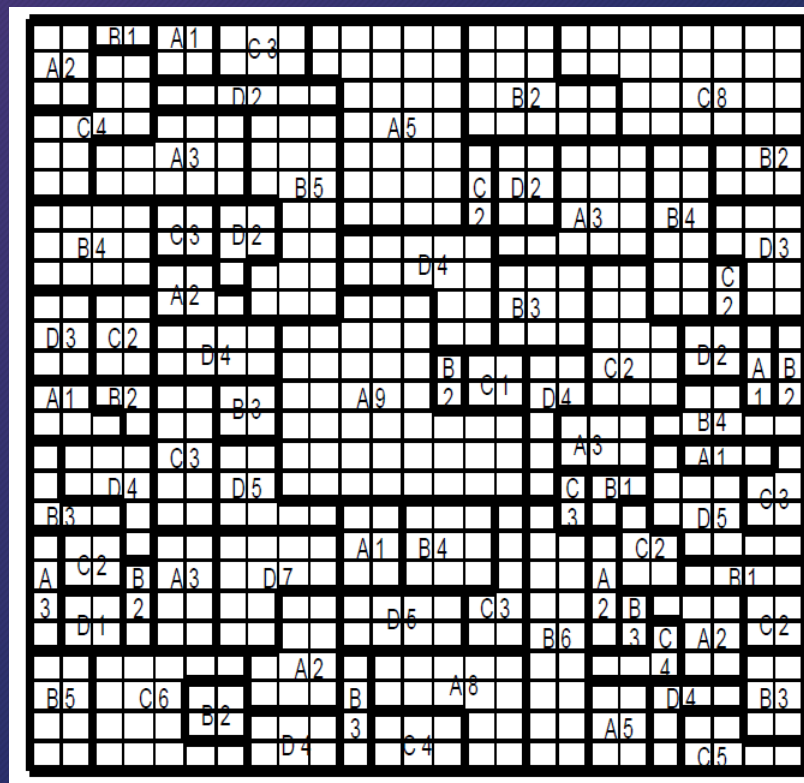
☝ Приметити да је бројилац обична Horvitz-Thompson-ова оцена, која је непристрасна тачкаста оцена популацијског тотала  $\tau_Y$ , али се и именилац може схватити као Horvitz-Thompson-ова оцена, која би била непристрасна оцена обима популације  $N$ .

Оцена тотала добија се из:  $\tau_Y = Nm_Y$ .

- Пример - узорак са неједнаким вероватноћама

Разматра се популација фарми, са припадним парцелама различитих површина и облика, чије су границе уцртане на квадратној мрежи димензија  $25 \times 25$ . Свака фарма означена је словом које представља тип фарме и бројем који представља број радника на фарми.

На следећој мапи дат је одговарајући графички приказ области од интереса.



Бира се узорак са понављањем обима  $n = 5$ .

Оцењује се: укупан број радника; укупан број фарми.

- Пример - узорак са неједнаким вероватноћама - наставак

Метод одабира узорка: узорковање се врши са вероватноћама које су пропорционалне површинама припадних парцела

✎ користи се додатна информација о јединицама у популацији

Имплементација метода одабира узорка:

претпостави се да је један квадратић мреже мерна јединица површине. Ако се на случајан начин (тј. са једнаком вероватноћом) одабере један квадратић на мрежи, и са  $x$  означи површина парцеле којој он припада, а са  $A = 625$  укупна површина области од интереса, онда је вероватноћа са којом је баш та парцела одабрана у узорак једнака:

$$\psi = \frac{x}{A} = \frac{x}{625}$$

✎ геометријске вероватноће

Случајан избор једног квадратића на мрежи изводи се тако што се узоркују по две целобројне вредности (између 1 и 25), које ће представљати његове координате.

Реализован узорак:

(11,18), (14,8), (5,7), (10,1), (25,14)

Оцењена вредност броја радника је:

- $\hat{\tau}_{\psi} = 310.0962$  Н-Н оцена
- $\hat{\tau}_{\pi} = 316.5735$  Н-Т оцена

Тачна вредност броја радника је:  $\tau_Y = 249$

Оцењена стандардна грешка оцене бр. радника једнака је:

- 60.99 за Н-Н оцену
- 41.87 за Н-Т оцену

Оцењена (Н-Т оцена) вредност броја фарми је: 98.44543

Тачна вредност броја фарми је: 79

