

Теорија узорака

7. април '20.

Стратификован узорак

- Стратификација (раслојавање) је подела популације на потпопулације, слојеве – **стратуме** (лат. 'stratum' – мн. 'strata')

Врши се класификација свих ентитета у популацији по стратумима, на основу неких додатних, претходно прикупљених / доступних информација. Као критеријум за стратификацију користи се нека (једна или више њих) карактеристика популације, за коју се сматра да је са посматраним обележјем у вези (односно да постоји известан вид зависности између ње и посматраног обележја од интереса).

Стратуми морају бити **међусобно дисјунктни**, тј. свака јединица популације може припадати тачно једном стратуму, и, при томе, стратуми морају „**покривати**“ **комплетну популацију**, тј. не сме се појавити јединица која није укључена ни у један стратум

- Након извршене стратификације узорци, унапред одређеног обима, бирају се из сваког стратума. При томе, одабири узорака из различитих стратума су међусобно независни и није неопходно користити исти метод узорковања за све стратуме. Описаним поступком добија се **стратификован узорак** ('stratified sample').

👉 Стратификован случајан узорак добија се одабиром
(простог) случајног узорка без понављања из сваког стратума

Зашто стратификован узорак?

- Потреба да се елиминише могућност одабира веома „лоших“ узорака (пре свега у смислу (не)репрезентативности)
- Прикупљање жељене количине података за подгрупе енитета у популацији (те подгрупе би тада требало сматрати стратумима) у циљу доношења закључака на нивоу подгрупа
- Дешава се да је метод стратификованог узорковања једноставнији за спровођење и да његова примена резултира мањим трошковима у односу на друге методе
- Повећање прецизности оцена непознатих популацијских параметара

Ознаке

- L - број стратума; N_h - број јединица у h -том стратуму, $h = \overline{1, L}$

$$N = \sum_{h=1}^L N_h \text{ — обим популације}$$

- n_h - обим узорка из h -тог стратума, $h = \overline{1, L}$

$$n = \sum_{h=1}^L n_h \text{ — (укупан) обим узорка}$$

унапред познате
(задате) вредности

- y_{hk} - вредност обележја Y јединице означене са k , која потиче из h -тог стратума
- вредности у вези са појединачним стратумом:

$$\tau_h = \sum_{k=1}^{N_h} y_{hk} \quad m_h = \frac{\tau_h}{N_h} \quad \sigma_h^2 = \frac{1}{N_h - 1} \sum_{k=1}^{N_h} (y_{hk} - m_h)^2$$

- популацијске вредности:

$$\tau_Y = \sum_{h=1}^L \tau_h \quad m_Y = \frac{\tau_Y}{N}$$

Општи теоријски резултати

Претпоставља се да је $\mathcal{S}_h$ вероватносни узорак из h -тог стратума, $h = \overline{1, L}$.

Приступ приликом закључивања заснован је на методу одабира узорка. Нека је са $\hat{t}_h$ означена непристрасна (у односу на метод) оцена τ_h , и нека је $D\hat{t}_h$ дисперзија ове тачкасте оцене. Кључни резултати у вези са непознатим популацијским тоталом τ_Y , односно непознатом популацијском средњом вредношћу m_Y , дати су у следећој табели:

тачкаста оцена $\hat{\theta}$ непознатог параметра θ	непристрасност	дисперзија $D\hat{\theta}$
$\hat{t}_Y^{str} = \sum_{h=1}^L \hat{t}_h$	✓	$\sum_{h=1}^L D\hat{t}_h$
$\hat{m}_Y^{str} = \frac{\hat{t}_Y^{str}}{N}$	✓	$\frac{1}{N^2} \sum_{h=1}^L D\hat{t}_h$

↑ дисперзије оцена за појединачне стратуме се сабирају, јер су одабири узорака из различитих стратума међусобно независни

Уколико се могу предложити непристрасне оцене $D\hat{t}_h$, $h = \overline{1, L}$, онда је непристрасна оцена $D\hat{t}_Y^{str}$ збир појединачних оцена дисперзије за сваки стратум, а непристрасна оцена $D\hat{m}_Y^{str}$ количник истог тог збира и N^2 .

Стратификован случајан узорак

Претпоставља се да је $\mathcal{S}_h$ (прост) случајан узорак без понављања из h -тог стратума, $h = \overline{1, L}$. Приступ приликом закључивања заснован је на методу одабира узорка. Кључни резултати у вези са непознатим популацијским тоталом τ_Y , односно непознатом популацијском средњом вредношћу m_Y , дати су у следећој табели:

тачкаста оцена $\hat{\theta}$ непознатог параметра θ	непристрасност	дисперзија $D\hat{\theta}$	оцена дисперзије $D\hat{\theta}$
$\hat{t}_Y^{str} = \sum_{h=1}^L N_h \cdot \bar{Y}_h$ $= \sum_{h=1}^L \sum_{k \in \mathcal{S}_h} \frac{N_h}{n_h} \cdot y_{hk}$	✓	$\sum_{h=1}^L N_h^2 \cdot \frac{\sigma_h^2}{n_h} \left(1 - \frac{n_h}{N_h}\right)$	$\sum_{h=1}^L N_h^2 \cdot \frac{\bar{S}_h^2}{n_h} \left(1 - \frac{n_h}{N_h}\right)$
$\hat{m}_Y^{str} = \frac{\hat{t}_Y^{str}}{N}$	✓	$\sum_{h=1}^L \left(\frac{N_h}{N}\right)^2 \cdot \frac{\sigma_h^2}{n_h} \left(1 - \frac{n_h}{N_h}\right)$	$\sum_{h=1}^L \left(\frac{N_h}{N}\right)^2 \cdot \frac{\bar{S}_h^2}{n_h} \left(1 - \frac{n_h}{N_h}\right)$

где су $\bar{Y}_h$ и $\bar{S}_h^2$, редом. узорачка средина и дисперзија обележја Y за h -ти стратум:

$$\bar{Y}_h = \frac{1}{n_h} \sum_{k \in \mathcal{S}_h} y_{hk} \quad \text{и} \quad \bar{S}_h^2 = \frac{1}{n_h - 1} \sum_{k \in \mathcal{S}_h} (y_{hk} - \bar{Y}_h)^2$$

Општи теоријски резултати

(наставак)

Кључни **асимптотски резултат** у вези са интервалном оценом непознатог популацијског тотала τ_Y , када је **приступ** приликом закључивања **заснован на методу** одабира узорка, је: апроксимативни $100 \cdot (1 - \alpha)\%$ двострани интервал поверења дат је са

$$\left[\hat{t}_Y^{str} - z_{1-\frac{\alpha}{2}} \sqrt{\widehat{D\hat{t}_Y^{str}}}, \hat{t}_Y^{str} + z_{1-\frac{\alpha}{2}} \sqrt{\widehat{D\hat{t}_Y^{str}}} \right]$$

где је $\widehat{D\hat{t}_Y^{str}}$ оцена $D\hat{t}_Y^{str}$, и то ако је задовољен тачно један од следећих услова:

- обим узорка из сваког појединачног стратума је „довољно велики“
- план узорковања је такав да постоји велики број стратума (са релативно малим обимом узорка из сваког појединачног стратума)

Уместо одговарајућег квантила $z_{1-\frac{\alpha}{2}}$ стандардне нормалне расподеле, под одређеним условима, може се користити $\left(1 - \frac{\alpha}{2}\right)$ -квантил Студентове расподеле са $(n - L)$ степени слободе, односно са апроксимативним бројем степени слободе (Satterthwaite's formula).

Резултати који се тичу непознате популацијске средње вредности m_Y лако се могу добити из резултата у вези са непознатим популацијским тоталом τ_Y , коришћењем везе:

$$\tau_Y = Nm_Y.$$

Планирање стратификованог узорковања

- Како дефинисати стратуме?
- Како одредити број стратума?
- Како распоредити узорак по појединачним стратумима?

Расподела / распоред обима узорка

Надаље се претпоставља да се ради са **стратификованим случајним узорком** обима n . Сматра се да су стратуми дефинисани и да је већ одређен и фиксиран обим узорка n , а приступа се одлучивању о обиму узорка n_h за сваки стратум појединачно, $h = \overline{1, L}$ ('sample size allocation').

У пракси се за решавање овог проблема обично користи нека од две популарне технике:

- пропорционални распоред
- Neyman-ов оптималан распоред

👉 Одређени распоред обима узорка по стратумима примењује се, пре свега, у циљу смањења дисперзије. Међутим, и други фактори могу условљавати расподелу обима узорка.

Главни принципи распореда приказани су на шеми десно:



Пропорционални распоред

Код пропорционалног распореда ('proportional allocation'), број јединица које се бирају у узорак из појединог стратума пропорционалан је броју јединица у том стратуму, тј:

$$n_h = n \cdot \frac{N_h}{N},$$

па је, дакле, стопа одабира узорка $\frac{n_h}{N_h}$ у сваком појединачном стратуму константна и једнака разломку узорковања за читаву популацију $\frac{n}{N}$.

За вероватноћу укључења првог реда важи:

$$\pi_{hk} = \frac{n}{N} \text{ за свако } k \text{ и свако } h.$$

👉 Приметити да је овде $\hat{m}_Y^{str}$ заправо узорачка средина (комплетног узорка)

Ова техника распоређивања обима стратификованог случајног узорка по стратумима резултира мањом дисперзијом одговарајућих оцена непознатих популацијских вредности у односу на (прост) случајан узорак без понављања:

- ако су обими стратума N_h „велики“ и / или
- ако се средње вредности m_h , $h = \overline{1, L}$, међусобно „веома разликују“

Оптималан распоред

Ако су дисперзије σ_h^2 мање-више једнаке у свим стратумима, пропорционални распоред је вероватно најбољи са становишта смањења дисперзија оцена.

Ако, међутим, постоје индиције да се појављују значајне разлике међу дисперзијама σ_h^2 у неким стратумима обично се примењује оптималан распоред ('optimum allocation'). Идеја је да расподела обима узорка по стратумима буде таква да се дисперзија оцене учини најмањом могућом.

- Неуман-ов распоред – минимизира дисперзију оцене за познат и фиксиран обим целог узорка n

$$n_h = n \cdot \frac{N_h \cdot \sigma_h}{\sum_{l=1}^L N_l \cdot \sigma_l}$$

👉 Тачне вредности стандардних одступања σ_h по стратумима обично нису познате; оне се онда оцењују коришћењем података из ранијих истраживања, постојећих сазнања или погађају на основу познавања неких других својстава јединица које чине стратум

Оптималан распоред (наставак)

- Оптималан распоред који минимизује дисперзију оцено за познат и фиксиран укупан трошак истраживања, односно минимизује укупан трошак истраживања за задату дисперзију оцено

Трошак обично укључује издатке за спровођење истраживања и прикупљање података, обрађивање и анализу података добијених истраживањем и сл. Најједноставнија функција којом се може описати трошак истраживања је линеарна функција облика:

$$C = C_0 + \sum_{h=1}^L c_h \cdot n_h ,$$

где је C_0 општи трошак а c_h трошак по јединици узоркованој из h -тог стратума. Ту се добија:

$$n_h = n \cdot \frac{\frac{N_h \cdot \sigma_h}{\sqrt{c_h}}}{\sum_{l=1}^L \frac{N_l \cdot \sigma_l}{\sqrt{c_l}}}$$

Оптималан распоред

(наставак)

Вредност n обима целог узорка, која минимизује дисперзију оцене за познат и фиксиран укупан трошак истраживања C , дата је формулом:

$$n = \frac{(C - C_0) \cdot \sum_{l=1}^L \frac{N_l \cdot \sigma_l}{\sqrt{c_l}}}{\sum_{l=1}^L N_l \cdot \sigma_l \sqrt{c_l}}$$

Ова техника распоређивања обима стратификованог случајног узорка по стратумима је таква да би већи узорак требало одабрати из стратума који је бројнији или унутар кога је варијабилност обележја Y већа, а узорак мањег обима из стратума у коме је скупље, односно теже спровођење узорковања.

Одређивање обима узорка

Различите технике распоређивања обима стратификованог случајног узорка по стратумима дају различите релативне обиме узорака $\frac{n_h}{n}$ по стратумима.

Сматра се да су стратуми дефинисани и да је установљен распоред, а задатак је одређивање обима узорка за оцењивање непознате популацијске средње вредности m_Y , тако да се постигне апсолутна грешка не већа од Δ са поверењем $1 - \alpha$. Користи се:

$$D\hat{m}_Y^{str} = \sum_{h=1}^L \left(\frac{N_h}{N}\right)^2 \cdot \frac{\sigma_h^2}{n_h} \left(1 - \frac{n_h}{N_h}\right) \leq \underbrace{\frac{1}{n} \sum_{h=1}^L \left(\frac{N_h}{N}\right)^2 \cdot \frac{n}{n_h} \cdot \sigma_h^2}_{v} = \frac{v}{n}$$

па, дакле, ако се фактори корекције због коначности стратума могу занемарити и испуњени су услови за примену нормалне апроксимације, онда је апроксимативни $100 \cdot (1 - \alpha)\%$ интервал поверења за m_Y дат са:

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

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

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

Дефинисање стратума

('stratification principle')

Из формула за дисперзије оцена непознатих популацијских вредности на основу стратификованог случајног узорка очигледно је да главни утицај на величину дисперзије оцене има утврђивање критеријума за стратификацију, а затим и одређивање броја (и, самим тим, величине) стратума.

Стратуме би требало формирати тако да их одликује **релативна хомогеност унутар самог стратума**, у смислу да вредности обележја Y , које је предмет истраживања, буду приближне на јединицама које чине стратум (σ_h^2 што мање за свако h).

Са друге стране, то значи да би требало инсистирати на **релативној хетерогености између стратума**, у смислу повећања диференцираности: средњих вредности обележја Y по стратумима (што је могуће веће разлике између m_h), односно стандардних девијација обележја Y по стратумима (што је могуће веће разлике између σ_h).

На број стратума утичу многобројни фактори. Генерално начело које би требало поштовати је: што је више релеватних сазнања о популацији расположиво - то би требало користити већи број стратума, али да, при томе, начин поделе популације на стратуме буде, што „природнији“.

👉 Обично је довољно 3-10 стратума за једнодимензионо обележје Y