

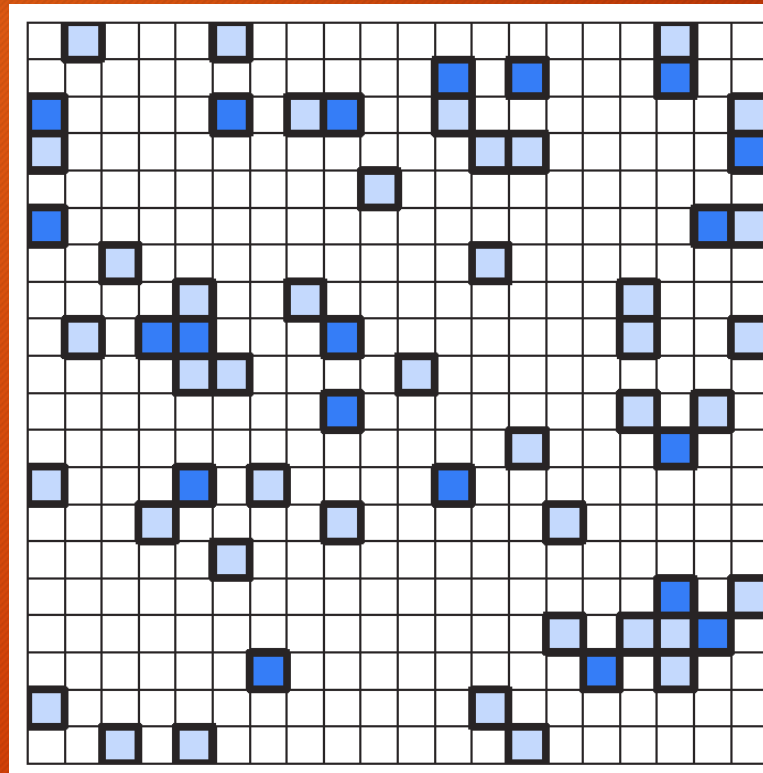
Теорија узорака

12. мај '20.

Двофазни узорак

- Раније је истицано да се помоћне информације, које се тичу додатних варијабилних својстава јединица посматрања у датој популацији, могу искористити како при пројектовању метода одабира узорка, тако и приликом закључивања, а све у циљу повећања прецизности оцена непознатих популацијских вредности главног обележја Y .
Тада је увек претпостављано да се **вредност помоћног обележја X** (које је, у општем случају, вишедимензионо) може одредити на произвољној јединици у популацији (у смислу, као да је унапред познато за СВАКИ ентитет). Међутим, у пракси је обично случај да је једини начин да се дође и до ових вредности – **узорковање**.
- Идеја **двофазног / двоструког узорковања** ('double / two-phase sampling') је да се у **фази I** одабере **почетни узорак** ентитета ради добијања искључиво помоћних информација, да би се, затим, у **фази II** одабрао други узорак ентитета на коме ће, поред вредности помоћног обележја, бити регистроване и вредности обележја од интереса. Тај други узорак обично је **подузорак почетног узорка**.

Двофазни узорак (наставак)



↑ Приметити:

- код двофазног узорка у свакој фази јединице узорковања су истог типа
- код двоетапног узорка у различитим етапама узоркују се јединице различитих типова (тј. различитих нивоа описане хијерархије)

Дат је једноставан приказ двофазног узорка.

- На јединицама обојеним тамно плаво регистроване су вредности и обележја X и обележја Y ,
- а на јединицама обојеним светло плаво регистроване су само вредности обележја X

Зашто двофазни узорак?

- Ако вредност помоћног обележја није унапред позната / доступна за произвољну јединицу посматрања, али је једноставније и / или јефтиније регистровати је на узоркованој јединици него што би то било случај са вредношћу главног обележја. Према томе, ако заиста постоји извешан вид зависности између обележја X и Y овај метод узорковања омогућава добијање „бољих“ оцена (у смислу прецизности) непознатих параметара популације за обележје Y .
- Поступање са неодговором.
- Узорковање јединица из ретких потпопулација.
- Побољшање оквира за одабир узорка (нпр. идентификовањем јединица које су иначе „тешко уочљиве“).

Ознаке

- N - обим популације (укупан број јединица посматрања, које чине популацију)
 - $n^{(1)}$ – обим (вероватносног) узорка $\mathcal{S}^{(1)}$ ентитета одабраних у фази I ('phase I sample') из комплетне популације; $n^{(2)}$ – обим (вероватносног) узорка $\mathcal{S}^{(2)}$ ентитета одабраних у фази II ('phase II sample') из узорка одабраног у фази I
- унапред познато

На јединицама из узорка $\mathcal{S}^{(1)}$ региструју се вредности обележја X , а на јединицама из узорка $\mathcal{S}^{(2)}$, додатно, и вредности обележја Y .

👉 Генерално гледано, може се сматрати да је обим узорка $\mathcal{S}^{(1)}$ релативно велики (то је подржано претпоставком да је испитивање јединица у циљу регистровања вредности помоћног обележја релативно једноставно и / или јефтино), те да може пружити тачну информацију о расподели вероватноћа обележја X

- y_k – вредност обележја Y ентитета означеног са k у оквиру за одабир узорка

Количничко оцењивање

Овде је помоћно обележје X варијабилна, **квантитативна карактеристика** ентитета за коју се сматра да је високо корелирана са главним обележјем Y . Претпоставља се да је $\mathcal{S}^{(1)}$ (прост) случајан узорак без понављања, а да је $\mathcal{S}^{(2)}$ подузорак $\mathcal{S}^{(1)}$ одабран, такође, применом метода SRSWOR. Са x_k означена је вредност обележја X ентитета означеног са k у оквиру за одабир узорка.

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

тачкаста оцена $\hat{\theta}$ непознатог параметра θ	непристрасност	апроксимација дисперзије $D\hat{\theta}$	оцена дисперзије $D\hat{\theta}$
$B := \frac{\tau_Y}{\tau_X}$ $\hat{B}^{(2)} = b^{(2)} = \frac{\sum_{k \in \mathcal{S}^{(2)}} y_k}{\sum_{k \in \mathcal{S}^{(2)}} x_k}$	✗		
$\hat{t}_X = \frac{N}{n^{(1)}} \sum_{k \in \mathcal{S}^{(1)}} x_k$	✓		
$\hat{t}_Y^{dr} = b^{(2)} \cdot \hat{t}_X$	✗	$N^2 \cdot \frac{\sigma_Y^2}{n^{(1)}} \left(1 - \frac{n^{(1)}}{N}\right) + N^2 \cdot \frac{\sigma_d^2}{n^{(2)}} \left(1 - \frac{n^{(2)}}{n^{(1)}}\right)$	$N^2 \cdot \frac{\bar{S}_Y^2}{n^{(1)}} \left(1 - \frac{n^{(1)}}{N}\right) + N^2 \cdot \frac{\bar{S}_e^2}{n^{(2)}} \left(1 - \frac{n^{(2)}}{n^{(1)}}\right)$

где су:

$$\sigma_d^2 := \frac{1}{N-1} \sum_{k=1}^N (y_k - Bx_k)^2; \bar{S}_Y^2 := \frac{1}{n^{(2)}-1} \sum_{k \in \mathcal{S}^{(2)}} (y_k - \bar{Y}^{(2)})^2, \bar{Y}^{(2)} := \frac{1}{n^{(2)}} \sum_{k \in \mathcal{S}^{(2)}} y_k \text{ и } \bar{S}_e^2 := \frac{1}{n^{(2)}-1} \sum_{k \in \mathcal{S}^{(2)}} (y_k - b^{(2)}x_k)^2$$

Планирање двофазног узорка за количничко оцењивање

- Колики би требало да буде обим узорка $\mathcal{S}^{(1)}$, односно његовог подузорка $\mathcal{S}^{(2)}$?

Идеалан однос обима узорка и његовог подузорка код двофазног узорковања, када је намера да се примени техника количничког оцењивања, зависи од односа трошка регистровања вредности помоћног обележја и трошка регистровања вредности главног обележја, као и од јачине линеарне везе (кроз координатни почетак) између та два обележја.

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

$$C = C^{(1)} \cdot n^{(1)} + C^{(2)} \cdot n^{(2)}$$

где је $C^{(1)}$ трошак регистровања вредности обележја X по ентитету и $C^{(2)}$ трошак регистровања вредности обележја Y по ентитету.

Разломак подузорковања, за који је минимизована дисперзија $D\hat{t}_Y^{dr}$ за познато и фиксирано C , дат је формулом:

$$\frac{n^{(2)}}{n^{(1)}} =: \vartheta = \sqrt{\frac{C^{(1)}}{C^{(2)}} \cdot \frac{\sigma_d^2}{\sigma_Y^2 - \sigma_d^2}}.$$

$$\begin{aligned} n^{(1)} &= \frac{C}{C^{(1)} + \vartheta C^{(2)}} \\ n^{(2)} &= \frac{C - n^{(1)} C^{(1)}}{C^{(2)}} \end{aligned}$$

Стратификација

Овде је помоћно обележје варијабилна, **квалитативна карактеристика** ентитета и користи се за дефинисање критеријума за стратификацију. Дакле, стратуми се одређују на основу помоћних информација добијених испитивањем узорка одабраног у фази I.

Претпоставља се да је $\mathcal{S}^{(1)}$ (прост) случајан узорак без понављања. Сваки ентитет из $\mathcal{S}^{(1)}$ класификован је у један, и тачно један од L стратума на основу познавања вредности

$$x_{kh} = \begin{cases} 1, & \text{ако јединица означена са } k \text{ припада } h - \text{ том стратуму} \\ 0, & \text{иначе} \end{cases}$$

регистроване за њега.

Гледано на нивоу популације, h -ти стратум сачињен је од N_h јединица, при чему је овде N_h непознато. Удео $\frac{N_h}{N}$ обима h -тог стратума у обиму популације N се, зато, мора

оцењивати и то узорачким уделом $\frac{n_h^{(1)}}{n^{(1)}}$, где је $n_h^{(1)}$ број јединица из $\mathcal{S}^{(1)}$ које припадају h -том стратуму.

Узорак $\mathcal{S}^{(2)}$ је стратификован случајан подузорак узорка $\mathcal{S}^{(1)}$. Наиме, из сваког стратума се од $n_h^{(1)}$ јединица из узорка $\mathcal{S}^{(1)}$, као прост случајан узорак без понављања, бира подузорак обима $n_h^{(2)}$ на коме се, затим, региструју вредности главног обележја Y .

☞ $n_h^{(2)}$ може зависити од фазе I узорковања

Стратификација (наставак)

Приступ приликом закључивања заснован је на методу одабира узорка. Кључни резултати у вези са непознатом популацијском средњом вредношћу m_Y , дати су у следећој табели:

тачкаста оцена $\hat{\theta}$ непознатог параметра θ	непристрасност	дисперзија $D\hat{\theta}$
$\hat{m}_Y^{dstr} = \sum_{h=1}^L \frac{n_h^{(1)}}{n^{(1)}} \cdot \bar{Y}_h^{(2)}$ $= \sum_{h=1}^L \sum_{k \in \mathcal{S}^{(2)}} \frac{n_h^{(1)}}{n^{(1)}} \cdot \frac{1}{n_h^{(2)}} \cdot x_{kh} y_k$	✓	$\frac{\sigma_Y^2}{n^{(1)}} \left(1 - \frac{n^{(1)}}{N}\right) + E \left(\sum_{h=1}^L \left(\frac{n_h^{(1)}}{n^{(1)}} \right)^2 \cdot \frac{\sigma_{h(1)}^2}{n_h^{(2)}} \left(1 - \frac{n_h^{(1)}}{n_h^{(2)}}\right) \right)$

други сабирак је остављен у облику мат. очекивања јер су :
 $n_h^{(1)}, n_h^{(2)}, \sigma_{h(1)}^2$
 случајне величине које зависе од узорка одабраног у фази I

где су:

$$\bar{Y}_h^{(2)} := \frac{1}{n_h^{(2)}} \sum_{k \in \mathcal{S}^{(2)}} x_{kh} y_k, \quad \sigma_{h(1)}^2 := \frac{1}{n_h^{(1)} - 1} \sum_{h=1}^L \sum_{k \in \mathcal{S}^{(1)}} x_{kh} \left(y_k - \bar{Y}_h^{(1)} \right)^2, \quad \bar{Y}_h^{(1)} := \frac{1}{n_h^{(1)}} \sum_{k \in \mathcal{S}^{(1)}} x_{kh} y_k.$$

Оцена дисперзије $D\hat{m}_Y^{dstr}$ дата је формулом:

$$\frac{N-1}{N} \sum_{h=1}^L \left(\frac{n_h^{(1)} - 1}{n^{(1)} - 1} - \frac{n_h^{(2)} - 1}{N - 1} \right) \frac{n_h^{(1)}}{n^{(1)}} \cdot \frac{\bar{S}_{h(2)}^2}{n_h^{(2)}} + \frac{1}{n^{(1)} - 1} \left(1 - \frac{n^{(1)}}{N} \right) \sum_{h=1}^L \frac{n_h^{(1)}}{n^{(1)}} \cdot \left(\bar{Y}_h^{(2)} - \hat{m}_Y^{dstr} \right)^2$$

где је

$$\bar{S}_{h(2)}^2 = \frac{1}{n_h^{(2)} - 1} \sum_{k \in \mathcal{S}^{(2)}} x_{kh} \left(y_k - \bar{Y}_h^{(2)} \right)^2.$$

- Пример - двофазни узорак

Оцењивање процента војних ветерана из тзв. вијетнамске ере (1961-1975. г) у болницама под надлежношћу U. S. Department of Veterans Affairs, који су заиста служили у Вијетнаму.

S. L. Lohr, “*Sampling: Design and Analysis*” (2010) - Example 12.1, Example 12.4

Грешке у истраживању

Укупна грешка резултата истраживања састоји се од две компоненте:

- ▷ узорачке грешке ('sampling error')
- ▷ неузорачке грешке ('nonsampling error')

које се морају засебно сагледати и одредити њихов квантитет.

Узорачка грешка	Неузорачка грешка
настаје услед самог узорковања, тј. зато што се врши делимично а не потпуно испитивање популације; њено постојање исказано је кроз разлику између оцењене и тачне (али непознате) вредности параметра популације за обележје од интереса	настаје услед несавршености процедура које се примењују у различитим фазама истраживања (може се појавити у било ком кораку планирања и спровођења истраживања на узорку или цензусу); најчешће је узрокована људским фактором
код вероватносног узорковања задаје се у вероватносним терминима	не може се приписати варијабилности која се појављује од узорка до узорка
одражава прецизност закључака	одражава валидност закључака
на њено смањење може се утицати повећањем обима узорка	не смањује се повећањем обима узорка, већ остаје присутна, а може чак бити и израженија у опсежнијим истраживањима
	много ју је теже мерити и контролисати

Неузорацке грешке

Неузорацке грешке, генерално гледано, представљају разне врсте пристрасности које настају у процесу истраживања. Разликује се више типова ових грешака, од којих су главне приказане на следећој шеми:



Неузорацке грешке (наставак)

- Грешка спецификације:
 - ◀ погрешна дефиниција (циљне) популације
 - ◀ разлика између стварних вероватноћа одабира јединица у узорак и теоријских вероватноћа одређених претпостављеним методом узорковања
- Грешка обухвата:
 - ◀ недоследно поступање анкетара, односно спроводилаца истраживања, са оквиrom за одабир узорка
 - ◀ дефектан оквир за одабир узорка - могући недостаци оквира су:
 - ▶ **грешка мањег обухвата** ('undercoverage') када у оквир нису укључене све јединице циљне популације
 - ▶ **грешка већег обухвата** ('overcoverage') када су у оквир укључене и јединице које не припадају циљној популацији
 - ▶ вишеструки уноси исте јединице у оквиру (грешка се појављује када се постојање мултипликата јединица не узме у обзир приликом анализе података)
- **Грешка мерења или одговора** - представља грешку која настаје када се регистрована вредност обележја од интереса (добијена мерењем извесним мерним инструментом или одговарањем на постављена питања) разликује од тачне (праве) вредности тог обележја (могући узроци: мерни уређај прави систематску грешку - пристрасан је или непрецизан; процес посматрања јединица у узорку утиче на регистроване вредности обележја; лоше формулисано питање у анкети, по правилу, резултира неважећим или погрешним одговором; под утицајем пристрасности анкетара испитаник „прилагођава“ свој одговор итд)

грешка неопсервације

грешке
опсервације

Неузорачке грешке (наставак)

- Грешка руковања подацима:
 - ненамерне грешке при регистровању / бележењу података
 - грешке при кодирању података
 - грешке при сређивању, односно табеларном / графичком приказивању података
- Грешка анализе:
 - пропусти у одабиру аналитичких техника (примена неадекватног апарата)
 - грешке при калибрацији података (нпр. невршење корекције тежина узорковања)
 - импутација - додељивање вредности недостајућим подацима (примењује се у случају неодговора, коришћењем заменске вредности, која нпр. потиче од јединице довољно „сличне“ посматраној, уместо вредности која недостаје)

👉 Постоји читав низ метода да се неузорачке грешке умање и ниједан да потпуно нестану. Важан задатак истраживача је да уложи сваки напор да одржи ове грешке што је могуће мањим.

Неодговор

Неодзив или неодговор односи се на ситуацију када је јединица посматрања узоркована, али вредности обележја од интереса за њу недостају (парцијално или комплетно).

Постоје два типа неодговора: **потпуни неодговор** ('unit nonresponse') - када за јединицу посматрања немамо никакву употребљиву прикупљену информацију и **делимични неодговор** ('item nonresponse') - када у информацијама прикупљеним за јединицу посматрања постоје „рупе“.

Уобичајени типови неодговора су:

- ◀ **недоступност** ентитета, тј. непостојање контакта се узоркованим ентитетом ('noncontact')
- ◀ **одбијање** ентитета да учествује у истраживању ('refusal')
- ◀ **неподобност** ентитета у погледу предмета истраживања ('ineligibility')

Грешка неодговора или неодзива је, дакле, грешка неопсервације и представља једну од најозбиљнијих неузорацких грешака. Императив истраживача требало би да буде управо спречавање, односно смањење неодговора, јер тенденција ка неодговору може бити повезана, на непознате начине, са обележјем које се проучава. Наиме, приликом појаве високог процента неодговора у узорку, природно је очекивати да преостале јединице у узорку (које су се одазвале истраживању) не буду репрезентативни представници целокупне популације. Неодговор ствара **пристрасност оцене** на чију величину директно утиче **степен различитости**, у погледу појаве која је предмет истраживања, између узоркованих ентитета за које су регистроване вредности обележја од интереса и узоркованих ентитета код којих је присутан неодговор, као и **удео неодговора** у читавом узорку.

Неодговор (наставак)

Најбоље би било да истраживач још у припремним фазама процеса истраживања учини све што је могуће како би спречио неодговор. Под тим се подразумева да добро упозна популацију и предвиди неке од потенцијалних узрока неодговора. То се може учинити ослањањем на ранија истраживања, односно спровођењем сопствених планираних експеримената и применом метода побољшања квалитета поступака прикупљања и обраде података.

У наставку су наведени неки фактори који могу имати утицаја на стопу (не)одговора и тачност прикупљених података:

- предмет истраживања (интересантност / актуелност / осетљивост теме и сл)
- временски период (доба дана, годишње доба и сл) када се спроводи истраживање
- анкетари (њихове квалификације, мотивација, обука, радна норма и сл)
- начин прикупљања података (поштом, електронским путем, телефоном, личним контактом)
- дизајн анкете
- оптерећење испитаника
- подстицаји \ дестимулације због учествовања \ неучествовања у истраживању
- накнадни, додатни контакт са ентитетом код кога је иницијално присутан неодзив

Повратни позиви и двофазни узорак

Повратни позиви ('callbacks') представљају покушаје поновног контактирања узоркованих јединица посматрања које претходно нису биле доступне или су одбиле да учествују у истраживању.

Један од првих метода за поступање са неодзивом предложили су Hansen и Hurwitz и он се састоји у подузорковању ентитета код којих је иницијално присутан неодзив и примена двофазног узорка за стратификацију у циљу оцењивања непознатих популацијских вредности обележја од интереса. Наиме, популација се подели на два стратума - један стратум чине ентитети који ће одмах пристати да учествују у истраживању и дају одговор, а други стратум чине ентитети који при првом контакту („позиву“) неће бити доступни или

ће одбити да учествују у истраживању. 👉 све до завршетка узорковања није познато ком стратуму припада појединачни ентитет

Затим се из тог другог стратума, сачињеног од неодговора, бира подузорак ентитета. При испитивању подузорка требало би уложити додатни напор, евентуално применити другачије, софистицираније (самим тим и скупље), добро осмишљене методе како би други позив био успешан и резултирао одговором. Из првог стратума, сачињеног од одговора, користе се **сви** ентитети из почетног узорка.

👉 двофазни узорак чине сви одговори из почетног узорка и подузорак неодговора из почетног узорка