Теорија узорака

21. април ’20.

ЛГлаваш

Постстратификација

До повећања прецизности оцена непознатих популацијских вредности може доћи коришћењем додатних података о популацији, на основу којих се врши стратификација.

Стратификација, међутим, није увек пожељан / погодан начин да се искористе подаци у вези са популацијом и то из неког од следећих разлога: постоји превише потенцијалних променљивих које би могле имати утицаја на дефинисање адекватног критеријума за стратификацију; за различите аспекте анализе, као најбољи се могу показати различити одабири стратума; за неке јединице је тешко одредити ком стратуму припадају и сл.

Тада се прибегава постстратификацији ('poststratification'), или тзв. „стратификацији након одабира узорка“ ('stratification after selection').

Постстратификација
(наставак)

Нека је обим посматране популације N. Бира се (прост) случајан узорак без понављања обима n из читаве популације.

Популација је подељена, према одређеној (помоћној) квалитативној карактеристици, на L дисјунктних група – постстратума. Приликом узорковања, за сваку јединицу која је одабрана у узорак забележена је (поред вредности главног обележја Y) и одговарајућа, реализована вредност уочене квалитативне карактеристике, чиме је омогућено њено класификовање у један од постстратума. При томе, сматра се да је унапред познат број јединица N_h у сваком постстратуму, $h = \overline{1, L}$. Нека је са y_{hk} означена вредност обележја Y јединице означене са k, која припада из h-том постстратуму.

ако $\frac{N_h}{N}$ није познато може се оценити применом двофазног узорка ('double sampling')

Број јединица n_h из комплетног узорка, за које је установљено да припадају h-том постстратуму, је случајна величина за свако $h = \overline{1, L}$.

Од раније је познато (презентација за 3. час, слајд 12) да је мат. очекивање сл. величине n_h једнако $\frac{nN_h}{N}$, па расподела узорка по постстратумима има тенденцију да апроксимира пропорционални распоред код стратификованог узорка. Кључни резултати у вези са непознатом популацијском средњом вредношћу m_Y дати су у следећој табели:

тачкаста оцена $\widehat{m}_Y^{post}$	$\sum_{h=1}^{L} \frac{N_h}{N} \cdot \bar{Y}_h$
апроксимација дисперзије $D\widehat{m}_Y^{post}$	$\frac{1}{nN}\left(1-\frac{n}{N}\right)\sum_{h=1}^{L} N_h\sigma_h^2 + \frac{1}{n^2}\left(1-\frac{n-1}{N-1}\right)\sum_{h=1}^{L}\left(1-\frac{N_h}{N}\right)\sigma_h^2$

валидна само при специјалним условима, (поред осталог, за „довољно велике“ узорке)

$\bar{Y}_h$ – узорачка средина
σ_h^2 – дисперзија обележја Y за h-ти постстратум

ЛГлаваш

- Пример – постстратификација

Наставак примера из презентације за 9. час, слајдови 2-5:

Следећи резултати добијени су на основу (простог) случајног узорка без понављања из API популације (apisrs) применом постстратификације приликом закључивања. Класификација школа одабраних у узорак извршена је по њиховом типу: E, M, H-школе.

Број узоркованих школа по постстратумима:

$$n_E = 142, \qquad n_M = 33, \qquad n_H = 25$$

Ако обележје Y представља број уписаних ученика у школи:

оцењена вредност популацијске средине је: $\widehat{m}_Y^{post} = 582.0567$
оцењена вредност стандардне грешке оцене $\widehat{m}_Y^{post}$ је: 19.99356

Ако обележје Y представља вредност API из 2000. г:

оцењена вредност популацијске средине је: $\widehat{m}_Y^{post} = 656.7816$
оцењена вредност стандардне грешке оцене $\widehat{m}_Y^{post}$ је: 9.196351

Групни узорак

Методи одабира узорка, код којих се појединачни ентитети бирају на случајан начин у узорак из целе популације или из стратума, на које је претходно подељена популација, погодни су за истраживања малих размера, али обично не и за велика и комплексна истраживања.

- Код (једноетапног) групног узорка ('single-stage cluster sampling'):
  јединица посматрања ≠ јединица узорковања
  па је, заправо, подела комплетне популације извршена на тзв:
  - примарне јединице ('primary sampling units') / групе / скупине / серије / кластери – јединице узорковања
  - секундарне јединице ('secondary sampling units') – јединице посматрања (ентитети)

- Након што се, по извесном принципу, изврши подела популације на више дисјунктних група, неким вероватносним методом узорковања одабере се одређени број група из којих се посматрају СВИ ентитети и тако формира групни узорак. Према томе, када се посматра једна група – или су све секундарне јединице од којих је она сачињена укључене у узорак или није ниједна.
- Дакле, иако ће вредности обележја Y суштински бити регистроване на секундарним јединицама из групног узорка, сам поступак узорковања врши се искључиво на скупу примарних јединица.

Зашто групни узорак?

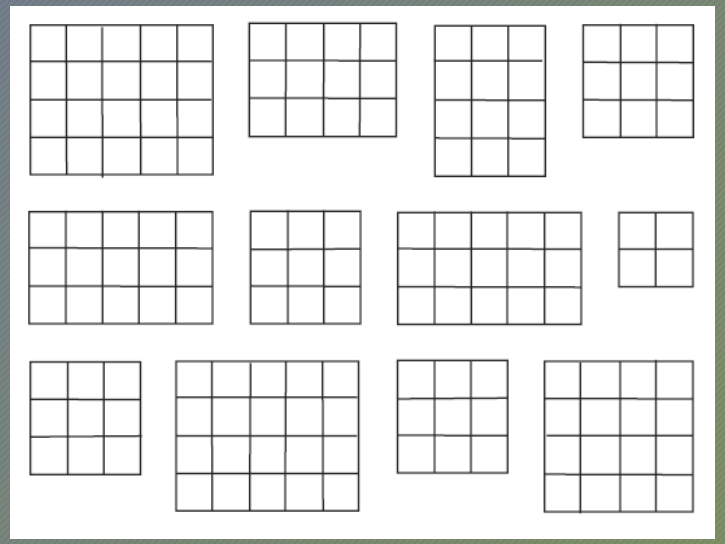
- Израда комплетног и дефинитивног оквира за одабир узорка може бити тешка, скупа или, чак, немогућа
- Популација може бити просторно врло широко распрострањена или већ може бити, на природан начин, организована у групе (нпр. градски стамбени блокови, домаћинства, одељења у школи итд). У таквим ситуацијама групни узорак обично је јефтинији од (простог) случајног узорка појединачних ентитета.

☝ Требало би, међутим, још напоменути да овде ентитети из исте групе имају тенденцију да буду међусобно сличнији него ентитети одабрани на случајан начин из комплетне популације. Ове сличности обично се јављају услед утицаја неких (најразличитијих) фактора који могу али не морају бити видљиви / мерљиви.

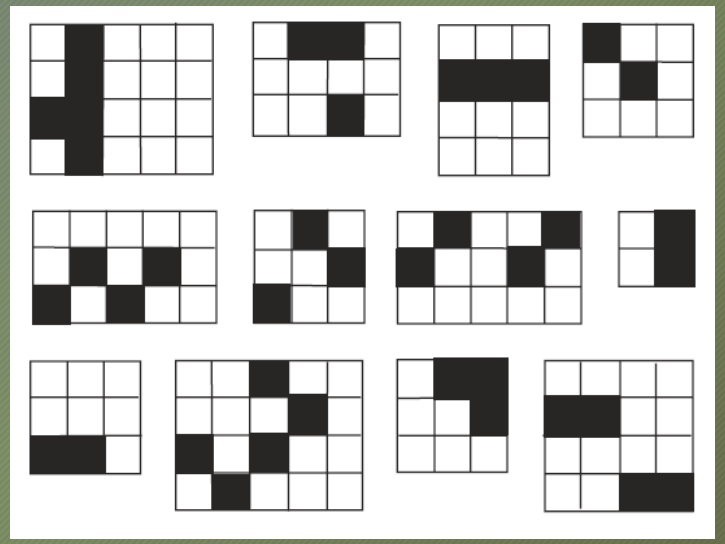
Испитивањем свих ентитета унутар узорковане групе, истраживач делимично понавља исту информацију, уместо добијања нових информација, што резултира смањењем прецизности оцена непознатих популацијских вредности

Групни  стратификовани узорак

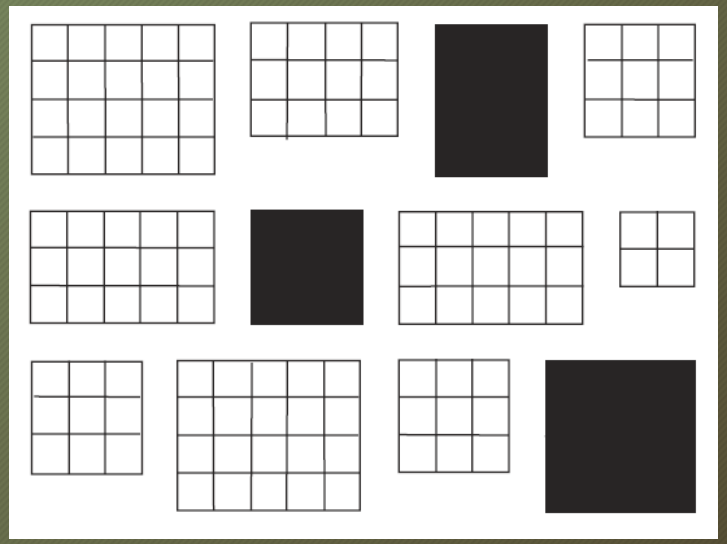
Популација подељена на:
L стратума, односно
N група



Одабир (простог) случајног узорка без понављања ентитета из сваког стратума



Одабир (простог) случајног узорка без понављања група – сви ентитети из одабраних група чине узорак



☝ Овакво директно упоређивање, као на горњим сликама, ипак, у општем случају, није могуће јер критеријум за поделу популације на групе може бити (и обично јесте) потпуно другачији од критеријума за стратификацију

ЛГлаваш

Ознаке

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

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

- n – обим узорка група, тј. примарних јединица
- y_{lk} – вредност обележја Y ентитета означеног са k, која потиче из l-те групе
- вредности у вези са појединачном групом:

$$\tau_l = \sum_{k=1}^{M_l} y_{lk} \qquad m_l = \frac{\tau_l}{M_l}$$

- популацијске вредности:

$$\tau_Y = \sum_{l=1}^{N} \tau_l \qquad \mu_Y = \frac{\tau_Y}{N} \qquad m_Y = \frac{\tau_Y}{M}$$

- $$\sigma_\tau^2 = \frac{1}{N-1} \sum_{l=1}^{N} (\tau_l - \mu_Y)^2 \qquad \sigma_l^2 = \frac{1}{M_l - 1} \sum_{k=1}^{M_l} (y_{lk} - m_l)^2$$

☝ унапред познато

SRSWOR метод одабира група

Претпоставља се да је $\mathcal{S}$ (прост) случајан узорак група, без понављања.

- 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	непристрасност	дисперзија $D\hat{\theta}$	оцена $\widehat{D\hat{\theta}}$ дисперзије $D\hat{\theta}$
$\hat{\tau}_Y^u = \frac{N}{n}\sum_{l\in\mathcal{S}} \tau_l = N \cdot \bar{Y}_\tau$	✓	$\frac{N^2}{n}\left(1-\frac{n}{N}\right)\sigma_\tau^2$	$\frac{N^2}{n}\left(1-\frac{n}{N}\right)\bar{S}_\tau^2$
$\hat{m}_Y^u = \frac{\hat{\tau}_Y^u}{M}$	✓	$\frac{D\hat{\tau}_Y^u}{M^2}$	$\frac{\widehat{D\hat{\tau}_Y^u}}{M^2}$
$\hat{\mu}_Y^u = \frac{\hat{\tau}_Y^u}{N}$	✓	$\frac{D\hat{\tau}_Y^u}{N^2}$	$\frac{\widehat{D\hat{\tau}_Y^u}}{N^2}$

$\pi_{lk} = \frac{n}{N}$

ℹ када је $M_l = \frac{M}{N}$, за $l = \overline{1,N}$, ово је, заправо, аритметичка средина m_l-ова за групе које су одабране у узорак

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

$$\bar{Y}_\tau = \frac{1}{n}\sum_{l\in\mathcal{S}} \tau_l \quad \text{и} \quad \bar{S}_\tau^2 = \frac{1}{n-1}\sum_{l\in\mathcal{S}} (\tau_l - \bar{Y}_\tau)^2$$

ЛГлаваш

SRSWOR метод одабира група
(наставак)

- Количничка оцена

Ако су тотали τ_l примарних јединица високо корелирани (ако постоји, корелација је обично позитивна) са одговарајућим „величинама“ M_l примарних јединица, онда се количничко оцењивање, где је помоћно обележје = величина групе, може показати ефикасним.

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

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

где је:

$$b := \frac{\sum_{l\in\mathcal{S}} \tau_l}{\sum_{l\in\mathcal{S}} M_l}$$

узорачки количник (популацијски количник B једнак је: m_Y)

ЛГлаваш

PPS метод одабира група

Претпоставља се да је одабран специјалан узорак (примарних јединица) са неједнаким вероватноћама – тзв. узорак група са вероватноћама пропорционалним „величини“ ('probability proportional to size sampling').

- Са понављањем- Hansen-Hurwitz-ова оцена

Случајан узорак $\mathcal{R}$ у наставку сматрамо неуређеним скупом кардиналности n, са евентуалним понављањима ознака група. За фиксирано l вероватноћа ψ_l одабира l-те групе дата је са:

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

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

тачкаста оцена $\widehat{\theta}$ непознатог параметра θ	непристрасност	оцена $\widehat{D\widehat{\theta}}$ дисперзије $D\widehat{\theta}$
$\hat{\tau}_Y^{\psi} = \frac{M}{n}\sum_{l\in\mathcal{R}}\frac{\tau_l}{M_l}$	✓	$\frac{M^2}{n(n-1)}\sum_{k\in\mathcal{R}}\left(m_l - \frac{\hat{\tau}_Y^{\psi}}{M}\right)^2$

непристрасна оцена m_Y

PPS метод одабира група
(наставак)

- Са понављањем – Horvitz-Thompson-ова оцена

Овде је $\mathcal{S}'$ редуковани узорак група са ефективним обимом узорка n_D.

Вероватноћа укључења првог реда π_l дата је формулом:

$$\pi_l = 1 - (1 - \psi_l)^n, \qquad \text{где је} \quad \psi_l = \frac{M_l}{M}.$$

Приступ приликом закључивања заснован је на методу одабира узорка. Кључни резултат је:

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

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

PPS узорковање

- Поменути план обично се у пракси имплементира неким од следећих поступака:
  - кумулативни метод ('cumulative-size method')
  - Lahiri-јев метод

  } за узорак са понављањем

  - 50+ поступака за узорак без понављања (K.R.W. Brewer and Muhammad Hanif, "*Sampling With Unequal Probabilities*", Springer Science+Business Media, LLC,1983)

☝ многи резултати овог типа настали су за потребе опсежних истраживања
у којима се прво изврши статификација (по одређеном критеријуму)
примарних јединица на велики број релативно малих стратума,
а затим из сваког стратума у узорак бира неколицина примарних јединица (најчешће две)

Планирање (једноетапног) групног узорковања

- Како дефинисати примарне јединице?
- Како одредити „величину“ примарне јединице?
- Колики би требало да буде обим узорка примарних јединица?

Дефинисање група

Како је група, овде, јединица узорковања – што је већи обим узорка група мање су дисперзије оцена непознатих популацијских вредности. Ове дисперзије зависе, пре свега, од варијабилности средњих вредности обележја Y по групама.

За постизање највеће прецизности, групе би требало формирати тако да их одликује релативна хомогеност између група, у смислу смањења диференцираности средњих вредности обележја Y по групама.

Са друге стране, то значи да би требало инсистирати на релативној хетерогености (секундарних јединица) унутар саме групе, у смислу да стандардна девијација обележја Y за појединачну групу буде што је могуће већа. Према томе, идеална примарна јединица била би она која у потпуности одражава разноликост ентитета у популацији, те је, као таква, „репрезентативна“.

Често је величина примарне јединице вредност која се сама намеће као природна. Када то није случај, генерално начело које би требало поштовати (бар у просторним истраживањима) је: што је већа група то је за очекивати већу варијабилност посматраног обележја унутар ње (⇨ повећање ефикасности групног узорка), при чему би требало водити рачуна и о следећем: подела популације на сувише велике групе може довести до смањења уштеде, која је типична за групни узорак и која представља једну од његових основних предности

Систематски узорак

- Систематско узорковање ('systematic sampling') у основи има исту структуру као групно узорковање.

- Код овог плана узорковања, појединачна примарна јединица (група) састоји се од секундарних јединица које су на известан систематски начин распоређене широм популације. У најједноставнијем контексту, код тзв. простог систематског / периодичног / механичког узорка, од овако формираних примарних јединица на случајан начин бира се једна.

- Нека је дата популација са N секундардних јединица (ентитета), које су у оквиру за одабир узорка (по одређеном редоследу) означене бројевима из скупа $\Omega = \{1, 2, \dots, N\}$ и нека постоје $n, K \in \mathbb{N}$, такви да важи: $N = n \cdot K$.
Бира се узорак ентитета обима n на следећи начин: од првих K ентитета из оквира одаберe се, (најчешће) на случајан начин, један ентитет као почетни, а затим сваки K-ти ентитет. Описаним поступком добија се периодични узорак са периодом / кораком K (K-ти линеарни систематски узорак). Нека је са y_k означена вредност обележја Y ентитета означеног са k.

☝ N, у општем случају, није целобројни умножак вредности K; тада се обими различитих узорака са периодом K могу разликовати за 1

Рбр. периодичног узорка		1	2	3	⋯	l	⋯	K
Структура узорка	1	y_1	y_2	y_3	⋯	y_l	⋯	y_K
	2	y_{K+1}	y_{K+2}	y_{K+3}	⋯	y_{K+l}	⋯	y_{2K}
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	n	$y_{(n-1)K+1}$	$y_{(n-1)K+2}$	$y_{(n-1)K+3}$	⋯	$y_{(n-1)K+l}$	⋯	y_N
Вероватноћа одабира узорка		$\frac{1}{K}$	$\frac{1}{K}$	$\frac{1}{K}$	⋯	$\frac{1}{K}$	⋯	$\frac{1}{K}$

Зашто систематски узорак?

- Одабир узорка је једноставнији (и мањи је број могућих узорака) него што је то случај код (простог) случајног узорка без понављања. Наиме, правило одабира секундардних јединица у узорак је сасвим просто, не захтева генерисање (псеудо)случајних бројева велики број пута нити потпуну нумерацију ентитета у популацији. Дакле, узорковање је често лакше за спровођење и то без пропратних грешака, као и економичније.

- Интуитивно је прихватљив – „равномерно“ је распоређен на комплетној популацији, не допушта случајна груписања или пропуштање заступљености неких делова популације (☝)

Нека је са $\mathcal{S}$ означен K-периодични узорак (његова реализација биће један од узорака из таблице на слајду 16).
Приступ приликом закључивања заснован је на методу одабира узорка.
Кључни резултати у вези са тачкастом оценом непознате популацијске средње вредности m_Y дати су у следећој табели (искоришћене су формуле за групни узорак обима 1 са слајда 9):

тачкаста оцена $\widehat{\boldsymbol{m}}_{\boldsymbol{Y}}^{\boldsymbol{sys}}$	$\bar{Y} = \frac{1}{n}\sum_{j\in\mathcal{S}} y_j$
$\boldsymbol{E}\widehat{\boldsymbol{m}}_{\boldsymbol{Y}}^{\boldsymbol{sys}}$	m_Y
$\boldsymbol{D}\widehat{\boldsymbol{m}}_{\boldsymbol{Y}}^{\boldsymbol{sys}}$	$\frac{1}{n^2K}\sum_{l=1}^{K}\left(\tau_l - \frac{\tau_Y}{K}\right)^2$
тачкаста оцена $\boldsymbol{D}\widehat{\boldsymbol{m}}_{\boldsymbol{Y}}^{\boldsymbol{sys}}$	???

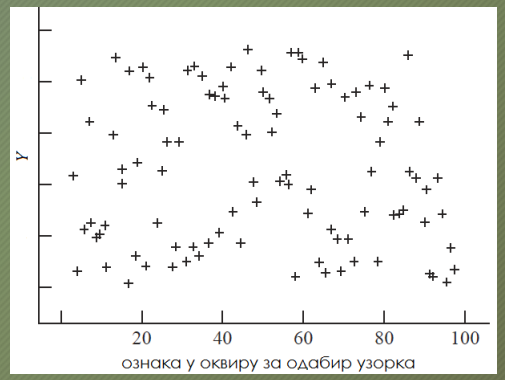
С обзиром да систематски узорак посматрамо као групни узорак обима 1 за добијање оцене $D\widehat{m}_Y^{sys}$ није могуће применити формулу са слајда 9 (вредност по формули је 0). Ту долази до изражаја главни проблем (и ограничење) у примени систематског узорка, а то је управо оцењивање дисперзија оцена непознатих популацијских вредности. Да би се овај проблем успешно решио потребно је имати додатне информације о структури популације.

Размотримо следећа три типа популације:

- ознаке у оквиру за одабир узорка придружене су ентитетима на случајан начин

  Овде се претпоставља да су ентитети на случајан начин распоређени у низ и затим означени, односно, суштински, да је посматрано обележје Y независно од распореда ентитета у оквиру за одабир узорка

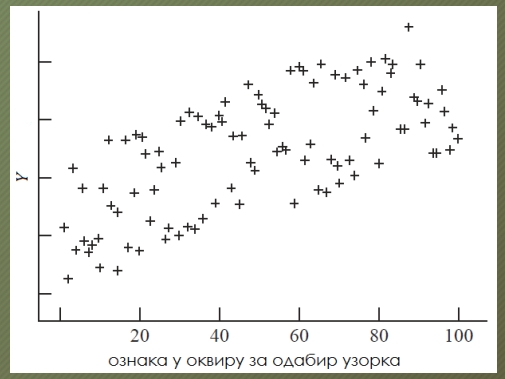
  Систематски узорак понаша као прост случајан узорак без понављања, па се могу примењивати исте формуле за оцене дисперзије $\widehat{D\bar{m}_Y^{sys}} = \frac{\bar{S}^2}{n}\left(1 - \frac{n}{N}\right)$, где је $\bar{S}^2$ узорачка дисперзија.



- оквир за одабир узорка уређен је монотоно (растуће или опадајуће)

  Посебан случај од интереса је постојање линеарног тренда зависности (тј. високе корелације) обележја Y од поретка ентитета у оквиру за одабир узорка.

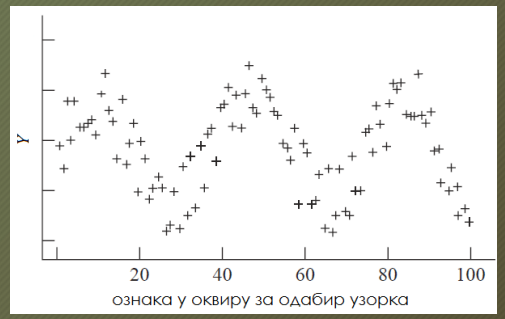
  Овде је систематски узорак знатно „бољи“ него прост случајан узорак без понављања, а (може бити) „лошији“ од стратификованог узорка кад су стратуми поднизови, дужине K, ентитета на узастопним местима у оквиру за одабир узорка.

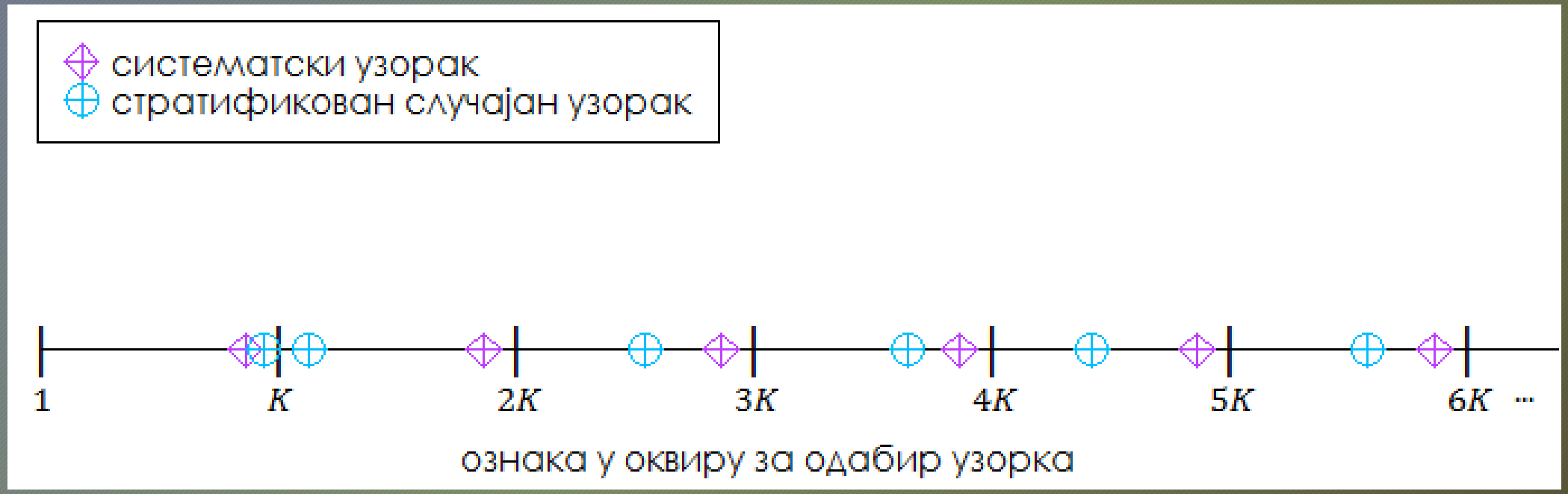


- изражен је периодичан шаблон у оквиру за одабир узорка

  Овде постоји периодичан тренд у низу вредности обележја Y, који одговара распореду ентитета у оквиру за одабир узорка.

  Ефикасност систематског узорка зависи од вредности периода K.




систематски узорак
стратификован случајан узорак
1
K
2K
3K
4K
5K
6K
...
ознака у оквиру за одабир узорка