

Теорија узорака

2. АПРИЛ '19.

► Пример – стратификован случајан узорак

Секретаријат за образовање савезне државе Калифорније је, у оквиру дела својих надлежности, спроводио годишње одређивање индекса академског успеха у државним школама у Калифорнији (Academic Performance Index). Овај индекс рачунат је на основу стандардизованих тестова, које су решавали ученици калифорнијских школа. Поред података о академским постигнућима ученика по школама, регистроване су и вредности различитих друштвено-економских обележја. API програм укинут је крајем 2018. г.

оквир – база података * – овде популацију чине све школе у Калифорнији са најмање 100 ученика, $N = 6194$; дати су компаративни подаци за период 1999-2000. г.

$n = 200$ је обим стратификованог случајног узорка ** – критеријум за стратификацију био је тип школе (тј. прецизније, ниво школовања (узраст):

E – elementary (5-12 г); M – middle (12-15 г); H – high (15-18 г) school)

обележје Y – број уписаних ученика у школи

* подаци се налазе у бази arjrop у оквиру  пакета survey

** узорак се налази у бази aristrat у истом пакету  Главаш

► Пример – стратификован случајан узорак – наставак

Стратификован случајан узорак распоређен је на следећи начин:

$$n_E = 100, \quad n_M = 50, \quad n_H = 50$$

Резон:

с обзиром на чињеницу да је уобичајено да Н-школе имају више ђака него М-школе, односно Е-школе, ако би се десило да је у (простом) случајном узорку без понављања одабран „већи“ број Н-школа то би водило ка „прецењивању“ средине и тотала обележја од интереса Y , док ако би се десило да случајан узорак садржи „мањи“ број Н-школа то би водило ка „потцењивању“ средине и тотала обележја Y . Фиксирање броја школа које се бирају из сваког појединачног нивоа требало би да има утицаја на смањење дисперзије.

$$\uparrow \quad N_E = 4421, \quad N_M = 1018, \quad N_H = 755$$

Оцењена вредност популацијске средине је: $\hat{m}_Y^{str} = 595.2821$

Оцењена вредност стандардне грешке оцено $\hat{m}_Y^{str}$ је: 18.50851

У истом пакету, у бази података `aprsrs`, налази се и (прост) случајан узорак без понављања изабран из API популације. На основу њега добијени су следећи резултати:

Оцењена вредност популацијске средине је: $\hat{m}_Y = \bar{Y} = 584.61$

Оцењена вредност стандардне грешке оцено $\hat{m}_Y$ је: 27.36837

► Пример – стратификован случајан узорак – наставак

Ако би обележје од интереса Y била вредност API из 2000. г, на основу описаног стратификованог случајног узорка, односно простог случајног узорка без понављања, добиле би се следеће реализације оцена:

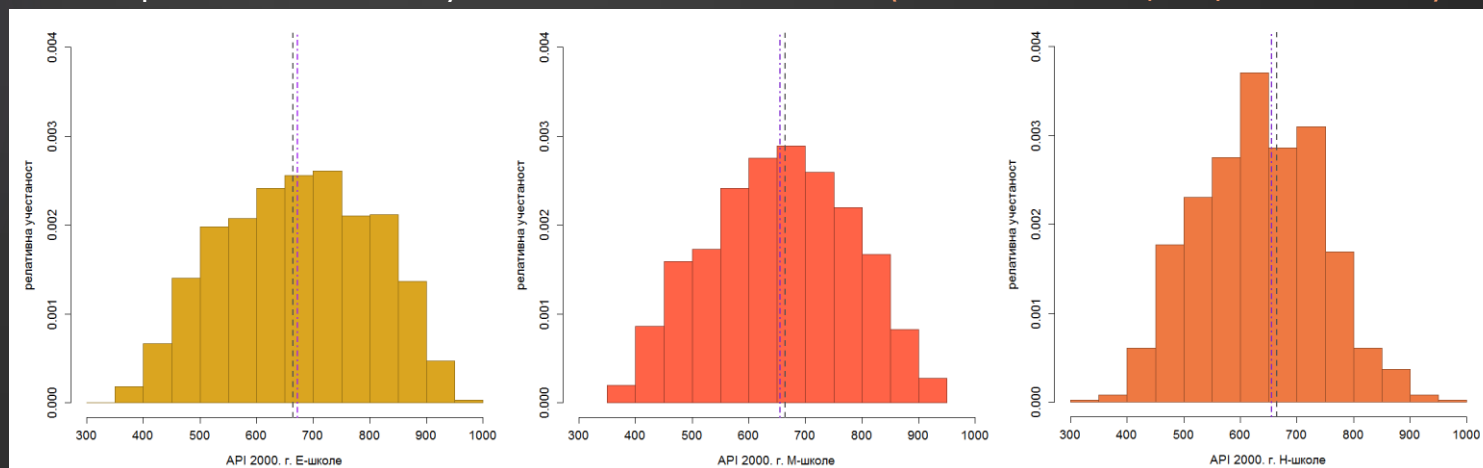
оцењена вредност популацијске средине: $\hat{m}_Y^{str} = 662.2874$
оцењена вредност стандардне грешке оцено $\hat{m}_Y^{str}$: 9.408941

$$m_Y = 664.7126$$

оцењена вредност популацијске средине: $\hat{m}_Y = \bar{Y} = 656.585$
оцењена вредност стандардне грешке оцено $\hat{m}_Y$: 9.249722

Одавде се види да применом стратификације, са задатим критеријумом за поделу на стратуме, не долази до значајног побољшања прецизности оцена непознатих популацијских вредности обележја Y . То је последица чињенице да стратуми нису довољно интерно хомогени ако се раслојавање врши на основу нивоа школовања (👉 посебно стратум Е-школа).

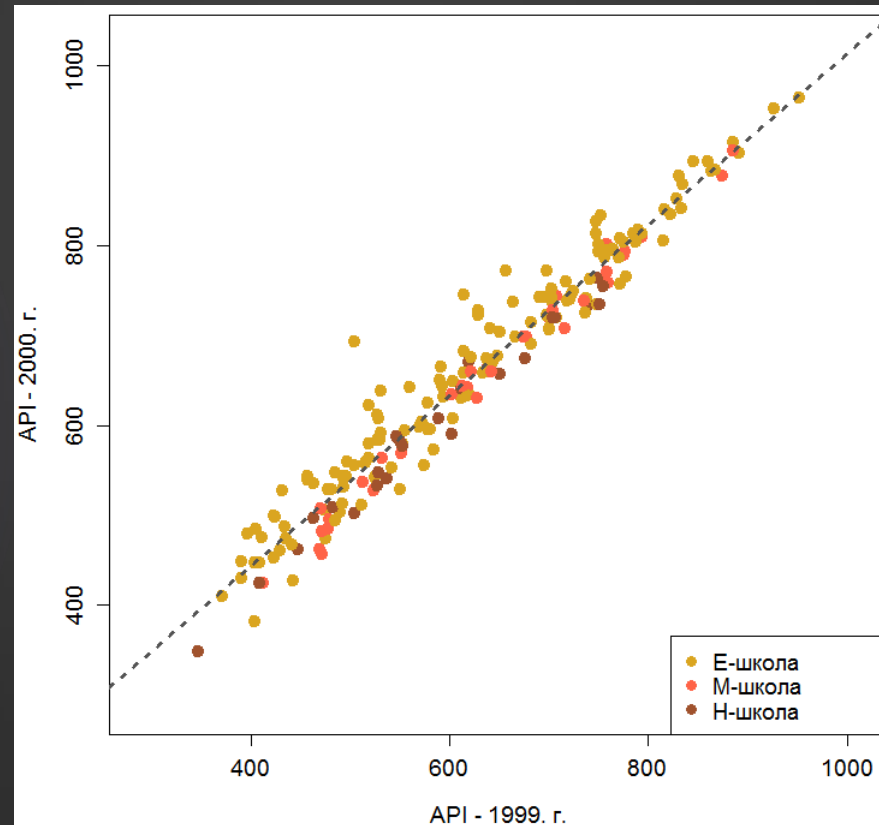
Хистограми вредности обележја Y из комплетне популације, приказани по стратумима.
Вертикалне праве:
популацијска средња вредност (сива)
средња вредност за стратум (љубичаста)



► Пример – стратификован случајан узорак – наставак

За обележје Y са претходног слајда (вредност API из 2000. г) помоћна информација (доступна за све калифорнијске школе), чије би коришћење заиста допринело побољшању прецизности оцене m_Y била би вредност API из претходне, 1999. г.

На следећој слици приказан је дијаграм распршености парова вредности главног и помоћног обележја, регистрованих на (простом) случајном узорку без понављања (apisrs):



► Пример – стратификован случајан узорак



Подручје на коме се врши истраживање је квадрат површине 40000 m^2 ($200 \text{ m} \times 200 \text{ m}$), смештен у једној старој шуми (Thomas County, Georgia, USA). Оно је представљено мрежом димензије 20×20 квадрата странице по 10 m . На доњој мапи дат је одговарајући графички приказ области од интереса, а уцртани бинарни подаци имају следеће значење:

1 := присуство, 0 := одсуство стабала дугоигличавог америчког бора унутар појединачног квадрата мреже.

Тачна вредност популацијске пропорције p , тј. удела квадрата мреже на којима расте ова врста бора, једнака је $p = \frac{249}{400} = 0.6225$.

Претпостави се да је истраживање вршено одабиром стратификованог случајног узорка без понављања обима $n = 40$ из четири стратума ($N_h = 100$, $h = \overline{1,4}$), као што је приказано на мапи, и то у складу са равномерним распоредом: $n_h = 10$, $h = \overline{1,4}$. Вредности у заградама су регистроване на квадратима одабраним у узорак.

1	1	1	1	(1)	1	1	(0)	0	0	(1)	1	0	1	0	1	1	1	0	1
1	1	1	0	1	0	0	0	1	1	1	1	0	1	1	1	0	1	0	1
1	1	1	1	1	1	0	(0)	0	1	1	1	1	1	1	1	1	1	(1)	1
0	1	1	0	0	0	(0)	(0)	(1)	1	1	1	1	1	1	1	1	1	1	0
1	1	0	1	1	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1
1	0	0	0	1	1	1	(0)	1	1	1	1	1	1	1	1	1	1	(1)	1
0	0	(1)	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	(1)	0
0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
0	0	(0)	0	1	0	1	1	1	(1)	1	1	1	1	1	1	1	1	1	0
1	1	1	1	1	0	1	0	1	0	1	0	1	1	1	1	(1)	1	1	1
0	1	1	0	(0)	1	(1)	1	(0)	0	1	1	1	1	1	1	1	1	1	1
0	0	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
0	1	(0)	0	1	(1)	0	1	1	1	1	1	1	1	1	1	1	1	(1)	1
1	0	0	1	1	0	(0)	0	(1)	1	1	1	1	1	1	1	1	1	(0)	0
0	(1)	0	1	1	0	1	1	0	1	1	1	1	1	1	1	1	1	(1)	1
0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	1	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	(0)	1	1	1	1	1	1	1	1	1	1	1	1	0
1	0	0	0	0	0	0	(0)	1	1	1	1	1	1	1	1	1	1	1	1

На основу стратификованог случајног узорка оцењује се популацијска пропорција p .

► Пример – стратификован случајан узорак – наставак

Тачкаста оцена (непознате) популацијске пропорције p је:

$$\hat{p}^{str} = \frac{1}{N} \cdot \sum_{h=1}^L N_h \cdot \frac{1}{n_h} \sum_{k \in \mathcal{S}_h} y_{hk} = \sum_{h=1}^L \frac{N_h}{N} \cdot \hat{p}_h^{str},$$

где је:

$$y_{hk} = \begin{cases} 1, & \text{ако у квадрату означеном са } k, \text{ из } h\text{-тог стратума, расте бор} \\ 0, & \text{иначе} \end{cases}.$$

Оцењена вредност популацијске пропорције је: $\hat{p}^{str} = \frac{3}{5} = 0.6$

Тачкаста оцена дисперзије ове оцене дата је формулом:

$$\sum_{h=1}^L \left(\frac{N_h}{N} \right)^2 \cdot \frac{\hat{p}_h^{str} (1 - \hat{p}_h^{str})}{n_h - 1} \left(1 - \frac{n_h}{N_h} \right)$$

(презентација за 3. час, слајд 6).

Оцењена вредност стандардне грешке оцене $\hat{p}^{str}$ је: **0.07071068**

Апроксимативни **95% интервал поверења** за популацијску пропорцију је: **[0.4565921, 0.7434079]**
(коришћен је 0.975-квантил Студентове расподеле са $36 (= n - L)$ степени слободе)

Стратификован случајан узорак

Предности	Мане
могућност оцењивања не само параметара целокупне популације него и параметара на нивоу подгрупа популације, тј. доношења закључака за појединачне стратуме и вршења поређења по стратумима	утврђивање релевантног критеријума за стратификацију може бити компликовано, посебно ако је истраживање само по себи комплексно и укључује велики број варијабилних својстава јединица
повећање репрезентативности узорка, јер омогућава заступљеност јединица из сваког стратума у крајњем (комплетном) узорку	неповољно је у ситуацијама када истраживач није у могућности да, на поуздан начин, класификује сваку јединицу у један и тачно један стратум; при томе, неопходно је да истраживач располаже комплетним и дефинитивним оквиром за одабир узорка
повећање прецизности оцена у смислу смањења њихових дисперзија (нпр. у односу на прост случајан узорак без понављања истог обима)	анализа података је сложен(иј)а
могућност да истраживач примењује различите методе узорковања од стратума до стратума, у зависности од сопствених потреба и доступности информација	врши се у складу са конкретним проблемом, уз претходно проучавање појаве и њене структуре. Стога, захтева велику количину претходних знања о популацији. Долазак до тих сазнања може представљати дуготрајан и скуп процес
може бити погодно за спровођење (али не увек и економично)	