

Теорија узорака

5. MART '19.

- Пример – одабир узорка коришћењем таблице случајних бројева
јединица узорковања – запослени у извесној компанији
оквир – платни списак
 $N = 200, n = 15$
обележје – број година радног стажа

36518	36777	89116	05542	29705	83775	21564	81639	27973	62413	85652	62817	57881
46132	81380	75635	19428	88048	08747	20092	12615	35046	67753	69630	10883	13683
31841	77367	40791	97402	27569	90184	02338	39318	54936	34641	95525	86316	87384
84180	93793	64953	51472	65358	23701	75230	47200	78176	85248	90589	74567	22633
78435	37586	07015	98729	76703	16224	97661	79907	06611	26501	93389	92725	68158
41859	94198	37182	61345	88857	53204	86721	59613	67494	17292	94457	89520	77771
13019	07274	51068	93129	40386	51731	44254	66685	72835	01270	42523	45323	63481
82448	72430	29041	59208	95266	33978	70958	60017	39723	00606	17956	19024	15819
25432	96593	83112	96997	55340	80312	78839	09815	16887	22228	06206	54272	83516
69226	38655	03811	08342	47863	02743	11547	38250	58140	98470	24364	99797	73498
25837	68821	66426	20496	84843	18360	91252	99134	48931	99538	21160	09411	44659

Посматра се делић једне таблице случајних бројева, приказан лево.

Коришћењем истог требало би одабрати узорак обима n из популације обима N . Ако је N k -тоцифрени број, из таблице се, почевши од било ког места „прочитају“, нпр. с лева на десно, k -тоцифрени бројеви и то њих тачно n , пропуштајући оне веће од N :

- ако је случајан узорак са понављањем
- и оне који су већ прочитани, ако је случајан узорак без понављања

365 **183** 677 789 **116 055** 422 970 583 775 215 648 **163** 927 973 624 **138** 565 262 817 578 814 613 281 380
756 351 942 888 **048 087** 472 **009** 212 615 350 466 775 369 630 **108** 831 368 331 841 773 674 **079 197** 402
275 699 **018** 402 338 393 **185** 493 634 641 955 258 631 687 384 841 809 379 364 953 514 726 535 823 701
752 304 720 **078 176** 🙄

Прочитани бројеви представљају ознаке јединица које су одабране у узорак

► Пример – (прост) случајан узорак



јединица узорковања – играч бејзбола са више од 100 излазака на палицу ("at bat") за своју екипу током САД бејзбол сезоне 2002. г.

оквир – званични подаци (Lahman's Baseball Database)*

$N = 438, n = 50$

обележје Y – број одиграних утакмица током бејзбол сезоне 2002. г.

Коришћењем програмског генератора добијен је следећи случајан узорак без понављања вредности посматраног обележја Y :

51 100 92 136 84 39 46 135 104 80 138 117 55 123 70 93 133 162 109 147 154 136 53 54 57 148
122 157 53 65 149 112 105 126 72 151 153 94 154 128 72 59 138 145 40 42 81 154 145 41

Оцењена вредност популацијске средње вредности је: $\hat{m}_Y = 103.48$

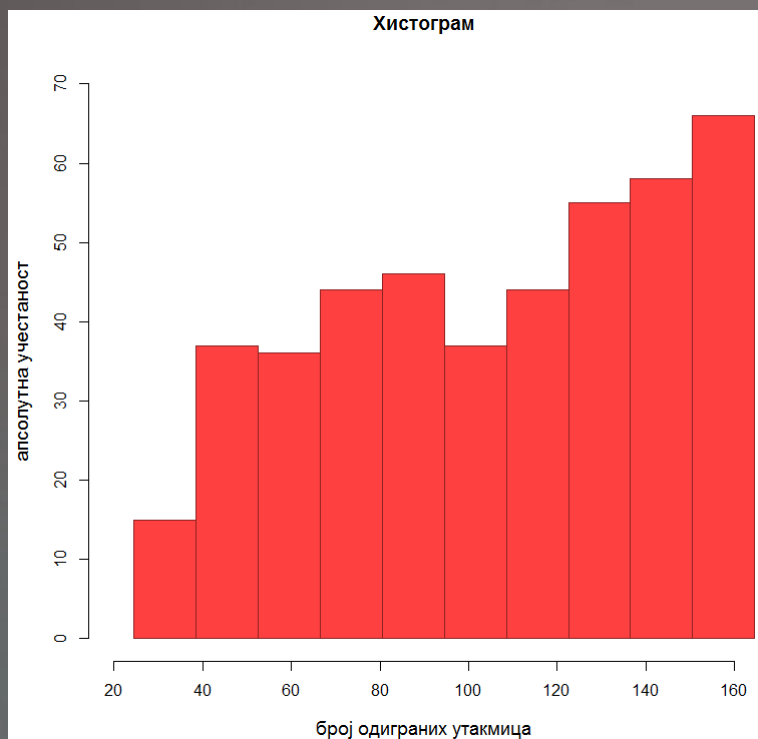
Тачна вредност популацијске средње вредности је: $m_Y = 105.3516$

Оцењена вредност дисперзије оцене $\hat{m}_Y$ једнака је: 28.51425

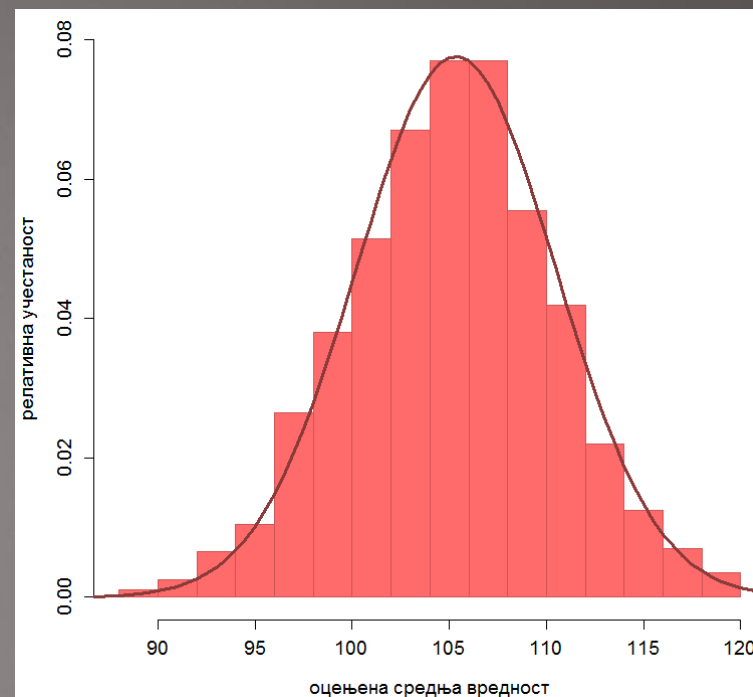
Оцењена стандардна грешка оцене $\hat{m}_Y$ једнака је: 5.339874

* подаци се налазе у бази batting у оквиру  пакета UsingR

► Пример – (прост) случајан узорак – наставак



хистограм апсолутне учестаности, којим је приказана расподела обележја Y на читавој популацији



хистограм релативне учестаности оцењених вредности популацијске средње вредности на основу 1000 случајних узорака без понављања, обима по 50

Апроксимативни 95% интервал поверења за популацијску средњу вредност је:
[92.74912, 114.2109] (коришћен је 0.975-квантил Студентове расподеле са 49 степени слободе)

SRSW(O)R

оцењивање пропорције

Претпостави се да је истраживач заинтересован за неко квалитативно својство јединица посматране популације

Популацијска пропорција p представља удео јединица популације које припадају посматраном нивоу категоричке променљиве од интереса, тј.

$$p = \frac{\text{број јединица у популацији које поседују дато својство}}{N}$$

Са статистичког становишта циљ је оценити непознати параметар p , при чему је обележје Y индикаторска случајна величина, дата са:

$$Y = \begin{cases} 1, & \text{ако је јединица поседује дато својство} \\ 0, & \text{иначе} \end{cases}$$

👉 За овако дефинисано обележје од интереса, популацијска средња вредност заправо је популацијска пропорција, па се могу користити сви кључни резултати у вези са оцењивањем непознате средње вредности

► Пример – (прост) случајан узорак

популација – све бебе рођене у току одређеног 24-часовног периода у
породилишту у Бризбејну, Аустралија

јединица узорковања – беба

оквир – болничка евиденција*

$N = 44, n = 8$

обележје Y – индикатор пола бебе, тј. $Y = \begin{cases} 1, & \text{ако је беба женског пола} \\ 0, & \text{ако је беба мушког пола} \end{cases}$

Коришћењем програмског генератора добијен је следећи случајан узорак **без понављања** вредности посматраног обележја Y :

1 0 0 1 1 1 1 0

Тачкаста оцена за непознату популацијску пропорцију p је (позната) узорачка пропорција

$$\hat{p} = \frac{1}{n} \sum_{k \in \mathcal{S}} y_k$$

Оцењена вредност популацијске пропорције је: $\hat{p} = \frac{5}{8}$

Тачна вредност популацијске пропорције је: $p = \frac{18}{44}$

Оцењена вредност дисперзије оцено $\hat{p}$ једнака је: 0.02739448

Формуле:

► узорачка дисперзија

$$\bar{S}^2 = \frac{n}{n-1} \hat{p}(1-\hat{p})$$

► тачкаста оцена $D\hat{p}$

$$\frac{\hat{p}(1-\hat{p})}{n-1} \left(1 - \frac{n}{N}\right)$$

* подаци се налазе у бази babyboom у оквиру  пакета UsingR

► Пример – (прост) случајан узорак – наставак

Егзактан интервал поверења за непознат популацијски тотал τ_Y (он овде представља укупан број новорођене деце женског пола), са поверењем бар $1 - \alpha$:

Идеја се базира на рачунању тачних вероватноћа из закона расподеле **хипергеометријске расподеле**. Наиме, узорачки тотал $T = \sum_{k \in S} y_k$, при задатој вредности популацијског тотала τ_Y , има хипергеометријску расподелу са параметрима N, τ_Y и n :

$$P\{T = l | \tau_Y\} = \frac{\binom{\tau_Y}{l} \binom{N - \tau_Y}{n - l}}{\binom{N}{n}}, \text{ за } l = \overline{0, n}$$

Нека је са a означена реализована вредност узорачког тотала T (она овде представља број женских беба у одабраном узорку). Границе $100 \cdot (1 - \alpha)\%$ интервала поверења $[\tau_L, \tau_U]$ за τ_Y одређују се на следећи начин:

- десна граница τ_U као највећи природан број такав да важи

$$P\{T \leq a | \tau_U\} > \alpha/2$$

- лева граница τ_L као најмањи природан број такав да важи

$$P\{T \geq a | \tau_L\} > \alpha/2$$

Егзактан интервал поверења са поверењем не мањим од 95% за популацијски тотал је: $[17, 39]$, а за популацијску пропорцију: $\left[\frac{17}{44}, \frac{39}{44}\right]$

► Пример – (прост) случајан узорак – наставак

Сада се претпостави да је, коришћењем програмског генератора, добијен следећи случајан узорак обима $n = 8$ са понављањем вредности посматраног обележја Y :

1 0 0 0 1 0 1 0

Егзактан интервал поверења за непознату популацијску пропорцију p , са поверењем бар $1 - \alpha$:

Идеја се базира на рачунању тачних вероватноћа из закона расподеле **биномне расподеле**. Наиме, узорачки тотал $T = \sum_{k \in S} y_k$, при задатој вредности популацијске пропорције p , има биномну $\mathcal{B}(n, p)$ расподелу.

Нека је са a означена реализована вредност узорачког тотала T . Ако је $a \neq 0$ и $a \neq n$, границе $100 \cdot (1 - \alpha)\%$ интервала поверења (p_L, p_U) за p одређују се на следећи начин:

► десна граница p_U као решење једначине

$$\sum_{l=0}^a \binom{n}{l} p_U^l (1 - p_U)^{n-l} = \alpha/2$$

► лева граница p_L као решење једначине

$$\sum_{l=0}^{n-a} \binom{n}{l} (1 - p_L)^l p_L^{n-l} = \alpha/2$$

Егзактан интервал поверења са поверењем не мањим од 95% за популацијску пропорцију је: $[0.08523341, 0.7551368]$

► Пример – (прост) случајан узорак

У децембру 2018. г, на националном нивоу, испитиван је став јавног мњења о увођењу стандарда за аутомобиле са циљем повећања енергетске ефикасности горива. У те сврхе спроведена је анкета на (простом) случајном узорку обима $n = 1012$ одраслих становника који поседују возачку дозволу. Њих $a = 810$ изјаснило се у корист увођења оваквих стандарда.

Нека је p непозната популацијска пропорција возача који подржавају увођење стандарда (она се може схватити и као (непозната) вероватноћа да (произвољан) возач подржава увођење стандарда).

👉 С обзиром на то да је истраживање вршено на националном нивоу, те да је, према томе, обим узорка занемарљиво мали у односу на обим популације, може се свуда изоставити фактор корекције $1 - \frac{n}{N}$

Одређује се апроксимативни 95% интервал поверења за популацијску пропорцију. То се врши на основу важења Централне граничне теореме, јер је n довољно велико:

$$P \left\{ \frac{|\hat{p} - p|}{\sqrt{p(1-p)}} \cdot \sqrt{n} \leq z_{1-\frac{\alpha}{2}} \right\} \approx 1 - \frac{\alpha}{2}$$

Овде важи:
$$D\hat{p} \approx \frac{p(1-p)}{n}$$

Облик интервала поверења је $[p_L, p_U]$, где су p_L и p_U решења квадратне једначине:

$$\left(n + z_{1-\frac{\alpha}{2}}^2\right) p^2 - \left(2n\hat{p} + z_{1-\frac{\alpha}{2}}^2\right) p + n\hat{p}^2 = 0$$

по променљивој p

Оцењена вредност популацијске пропорције је: $\hat{p} = \frac{810}{1012}$

Апроксимативни 95% интервал поверења за популацијску пропорцију је: $[0.775, 0.824]$

► Пример – (прост) случајан узорак

Планирано је истраживање јавног здравља на широј територији једног великог града ради процене удела деце, узраста 0-14 година, на којој није спроведена имунизација (вакцинација) против дечје парализе. Организатори овог пројекта желели би да постигну да узорачки удео неадекватно имунизоване деце са вероватноћом бар 98% одступа, по апсолутној вредности, од тачног удела p за највише 0.05. Одредити колики би требало да буде обим узорка.

👉 С обзиром на претпоставке и овде се може изоставити фактор корекције и применити нормална апроксимација

Формула за одређивање траженог обима узорка своди се на:

$$n_0 = \frac{p(1-p)z_{1-\frac{\alpha}{2}}^2}{\Delta^2},$$

где је апсолутна грешка $\Delta = 0.05$.

Приметити да наведена формула зависи од непознате популацијске пропорције p .

Најгрубља могућа процена за p је $p = 0.5$, па је тада најмањи обим узорка, за задату тачност, дат вредношћу израза:

$$\frac{z_{1-\frac{\alpha}{2}}^2}{4\Delta^2}$$

Обим узорка за оцењивање непознате популацијске пропорције је: 542

► Пример – (прост) случајан узорак – наставак

☝ Ако постоји оправдан разлог због кога се верује да је p обавезно мање од неког броја r_1 , $r_1 < 0.5$, односно веће од неког броја r_2 , $r_2 > 0.5$, фактор $p(1 - p)$ у формули за израчунавање обима узорка n , може да се замени са $r_1(1 - r_1)$, односно са $r_2(1 - r_2)$, што има утицај на (понекад и значајно) смањење обима узорка потребног за оцењивање p са задатом тачношћу

Ако нпр. претходна истраживања указују да је не више од 20% деце посматраног узраста неадекватно имунизовано добија се вредност 347 за обим узорка (редукција обима узорка за око 36%)

SRSWOR – оцењивање параметара потпопулације

Нека је обим посматране популације N и нека је N_l број јединица те популације које припадају њеној l -тој потпопулацији. Нека је са y_{lk} означена вредност обележја Y ентитета означеног са k , који потиче из l -те потпопулације.

Потпопулацијски тотал, односно средња вредност, дати су са:

$$\tau_l = \sum_{k=1}^{N_l} y_{lk}, \quad m_l = \frac{1}{N_l} \sum_{k=1}^{N_l} y_{lk}$$

Бира се узорак без понављања обима n из читаве популације. Тачно n_l ентитета (од одабраних n) потиче из l -те потпопулације и они чине подузорак $\mathcal{S}_l$.

тачкаста оцена $\hat{m}_l$	$\frac{1}{n_l} \sum_{k \in \mathcal{S}_l} y_{lk}$
$E\hat{m}_l$	m_l
$D\hat{m}_l$	$\sigma_l^2 \left(E \left(\frac{1}{n_l} \right) - \frac{1}{N_l} \right)$
тачкаста оцена $D\hat{m}_l$	$\frac{\bar{S}_l^2}{n_l} \left(1 - \frac{n_l}{N_l} \right)$

када је обим N_l l -те потпопулације непознат (што је често случај),
фактор корекције због коначности популације $1 - \frac{n_l}{N_l}$, замењује се својим математичким очекивањем $1 - \frac{n}{N}$

σ_l^2 (непозната) потпопулацијска дисперзија за јединице из l -те потпопулације,
а $\bar{S}_l^2$ (позната) узорачка дисперзија за јединице из подузорка $\mathcal{S}_l$

☝ Ако је познат обим N_l l -те потпопулације резултати који се тичу непознатог потпопулацијског тотала лако се добијају из оних за потпопулацијску средњу вредност m_l коришћењем везе: $\tau_l = N_l m_l$.

Ако није познат обим N_l као непристрасна оцена за τ_l може се користити статистика:

$$\hat{\tau}_l = \frac{N}{n} \sum_{k \in \mathcal{S}_l} y_{lk}$$

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 sukcesивно – 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 π_k , односно ψ_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 ψ_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 τ_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 $\hat{t}_\psi$

$$\frac{1}{n} \sum_{k \in \mathcal{R}} \frac{y_k}{\psi_k}$$

Hansen-Hurwitz-ова оцена

$$E\hat{t}_\psi$$

$$\tau_Y$$

тачкаста
оцена $D\hat{t}_\psi$

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

СА / БЕЗ ПОНАВЉАЊА

Нека су познате вероватноће укључења првог реда π_k и вероватноће укључења другог реда π_{kl} ; за све њих претпоставља се да су строго позитивне.

Кључни резултати у вези непознатим популацијским тоталом τ_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 $\hat{t}_\pi$	$\sum_{k \in \mathcal{S}'} \frac{y_k}{\pi_k}$	Horvitz-Thompson-ова оцена
$E \hat{t}_\pi$	τ_Y	
тачкаста оцена $D \hat{t}_\pi$	$\sum_{k \in \mathcal{S}'} \left(\frac{1}{\pi_k^2} - \frac{1}{\pi_k} \right) y_k^2 + \sum_{k \in \mathcal{S}'} \sum_{\substack{l \in \mathcal{S}' \\ l \neq k}} \left(\frac{1}{\pi_k \pi_l} - \frac{1}{\pi_{kl}} \right) y_k y_l$	→ 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 τ_Y дати су у следећој табели:

тачкаста оцена $\hat{t}_\pi$	$\sum_{k \in \mathcal{S}} \frac{y_k}{\pi_k}$
$E \hat{t}_\pi$	τ_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 τ_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. π_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 τ_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: $\tau_Y = Nm_Y$.