ТЕОРИЈА УЗОРАКА час 6

9. април ’14.

УЗОРАК СА НЕЈЕДНАКИМ ВЕРОВАТНОЋАМА (*Unequal probability Sampling*)

- Узорак са неједнаким вероватноћама често се користи у пракси, док прост случајан узорак (код кога свака јединица популације има подједнаку вероватноћу да буде изабрана у узорак) има, углавном, теоријски значај.
- Прост случајан узорак не узима у обзир инхерентно одступање у вредностима обележја јединица популације. Стога ће ова стратегија вероватно дати резултате који нису у потпуности поуздани, посебно када је то одступање значајно.

- Код популација код којих је управо таква ситуација може се прибећи другим плановима узорковања, под условом да су доступне додатне информације о одговарајућој променљивој за СВЕ јединице популације.
- Таква променљива назива се величина (*'size' variable*).
- Примери:
  - врши се истраживање на нивоу државе, региони не морају бити подједнако важни
  - ако је јединица популације породица, величина је број чланова породице
  - ако је фирма јединица популације, величина је нпр. број запослених, биланс пословања и сл.
- Поступак избора елемената у узорак са вероватноћом пропорционалном величини може бити са понављањем или без понављања.

УЗОРАК СА ВЕРОВАТНОЋОМ ПРОПОРЦИОНАЛНОМ ВЕЛИЧИНИ СА ПОНАВЉАЊЕМ

(*Probability Proportional to Size With Replacement Sampling Method*)

- Нека су X_i и Y_i, редом, вредност променљиве која представља величину и вредност обележја i-те јединице, $i = 1, 2, \dots, N$. Претпоставља се да су свих N вредности $X_1, X_2, \dots, X_N$ познате.
- Узорак обима n се добија извлачењем, n пута са понављањем, јединица популације, при чему је у сваком извлачењу вероватноћа избора i-те јединице (*selection probability*), у ознаци P_i, пропорционална њеној величини.

  Јасно је да је $P_i = \frac{X_i}{X}$, $i = 1, 2, \dots, N$, где је $X = \sum_{i=1}^{N} X_i$.

ПОСТУПЦИ ЗА ИЗБОР УЗОРАКА

- *Cumulative total method*

Формирају се кумуланте:

$$T_1 = X_1,\ T_2 = X_1 + X_2,\ \ldots,\ T_N = X_1 + X_2 + \cdots + X_N$$

Изабере се случајан број R између 1 и X. Ако је $T_{i-1} < R \leq T_i$ у узорак се бира i-та јединица ($T_0 = 0$). Ова процедура понавља се n пута, све док се не добије узорак обима n.

Описани метод је врло тешко имплементирати код популација великог обима.

- *Lahiri's method*

Нека је $M = \max_{i=1,2,\ldots,N} X_i$.

КОРАК I Одабери случајан број i између 1 и N

КОРАК II Одабери случајан број R између 1 и M

Ако је $R \leq X_i$, у узорак се бира i-та јединица. Иначе се она одбацује и понављају кораци I и II док се не одабере (нека) јединица.

Hansen-Hurwitz-ова оцена (*H – H Estimator*)

- Оцена тотала:

  Нека је y_i вредност обележја y јединице одабране у узорак, у i-том извлачењу, а p_i одговарајућа вероватноћа избора те јединице, $i = 1, 2, \dots, n$.

  Непристрасна оцена тотала обележја популације је $\hat{Y}_{H-H} = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{p_i}$, а њена дисперзија је

$$V[\hat{Y}_{H-H}] = \frac{1}{n}\sum_{i=1}^{N} P_i\left(\frac{Y_i}{P_i} - Y\right)^2$$

- Непристрасна оцена $V[\hat{Y}_{H-H}]$ је

$$v[\hat{Y}_{H-H}] = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\frac{y_i}{p_i} - \hat{Y}_{H-H}\right)^2$$

- Оцена средине:

  Ако је $\hat{\bar{Y}} = \frac{\hat{Y}_{H-H}}{N}$, онда је $\hat{\bar{Y}}$ непристрасна оцена средине обележја популације, а непристрасна оцена њене дисперзије је $v\left[\hat{\bar{Y}}\right] = \frac{1}{n(n-1)} \sum_{i=1}^{n} \left(\frac{y_i}{Np_i} - \hat{\bar{Y}}\right)^2$

- Предности:
  - довољно је знати вероватноће избора p_i само за јединице популације, које су одабране у узорак
  - за оцену тотала обележја популације није потребно знати обим популације

УЗОРАК СА ВЕРОВАТНОЋОМ ПРОПОРЦИОНАЛНОМ ВЕЛИЧИНИ БЕЗ ПОНАВЉАЊА

(*Probability Proportional to Size WithOut Replacement Sampling Method*)

- Узорак обима n се добија извлачењем, n пута без понављања, јединица популације, при чему се избор јединице врши са вероватноћом пропорционалном величини. Вероватноће избора се мењају у сваком (од њих n) извлачењу.
- Стога се природно појављују друге оцене непознатих параметара, које узимају у обзир овај проблем.
- *Desraj ordered estimator*
- *Murthy's ordered estimator*
- *Horvitz-Thompson estimator*

Horvitz-Thompson-ова оцена (*H − T Estimator*)

- То је општа оцена за тотал популације, која се може примењивати на било који вероватносну стратегију узорковања. Могуће је користити је како за узорке без понављања, тако и за узорке са понављањем.
- Ова оцена базира се на познавању вероватноћа укључења првог реда.
- Захтева се да вероватноће укључења првог и другог реда буду (строго) позитивне.

- Оцена тотала за узорак без понављања:

  Непристрасна оцена тотала обележја популације је $\hat{Y}_{H-T} = \sum_{i=1}^{N} \frac{Y_i I_i}{\pi_i}$, а непристрасна оцена њене дисперзије је $v\left[\hat{Y}_{H-T}\right] = \sum_{i=1}^{N}\sum_{j=1}^{N} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j \pi_{ij}} Y_i Y_j I_i I_j$.

- Оцена тотала за узорак са понављањем:

  Прво би требало направити редуковани узорак s' (тако што се из почетног узорка избаце се јединице за које постоје вишеструка понављања); формуле су:

$$\hat{Y}_{H-T} = \sum_{i \in s'} \frac{Y_i}{\pi_i}$$

$$v\left[\hat{Y}_{H-T}\right] = \sum_{i \in s'} \left(\frac{1}{\pi_i^2} - \frac{1}{\pi_i}\right) Y_i^2 + 2 \sum_{i \in s'} \sum_{\substack{j \in s' \\ i<j}} \left(\frac{1}{\pi_i \pi_j} - \frac{1}{\pi_{ij}}\right) Y_i Y_j$$

Требало би приметити да претходни израз за оцену дисперзије може бити негативан. Постоји алтернативна оцена, која је, при важењу довољних услова, увек ненегативна и сматра се стабилнијом (*Sen-Yates-Grundy*).

- Код *PPSWOR* узорка:

  Може се користити Horvitz-Thompson-ова оцена, под претпоставком да су вероватноће укључења познате. Међутим, код овог узорка, генерално гледано, нису доступни експлицитни изрази за вероватноће укључења. Уз помоћ рачунара, истраживач може направити списак свих могућих резултата извлачења n јединица у узорак и срачунати вероватноће укључења.

- Код *PPSWR* узорка:

$$\pi_i = 1 - (1 - p_i)^n \quad \pi_{ij} = \pi_i + \pi_j - [1 - (1 - p_i - p_j)^n]$$

- Функција `sample()` наставак

  `sample(x, size, replace=F, prob=NULL)`

  Аргументом `prob` могу се назначити различите вероватноће, са којима се узоркују појединачни елементи из датог вектора `x`. У пракси то се спроводи постављањем овог аргумента на вектор "тежина" са којима се узоркује сваки елемент из вектора `x`. Вектор "тежина" мора бити исте дужине као и дати вектор `x`, а његови елементи морају бити ненегативни бројеви, не смеју сви бити једнаки нули, и не морају бити нормирани јединицом.

```
> sample(c("P", "G"), 10, replace=T, prob=c(.9, .1))
 [1] "G" "P" "G" "P" "P" "P" "P" "P" "P" "P"
> sample(c("P", "G"), 10, replace=T, prob=c(.9, .1))
 [1] "P" "P" "P" "P" "P" "P" "P" "P" "P" "P"
> p <- c(1, 2, 3, 4, 5, 5, 4, 3, 2, 1)
> x <- 1:10
> sapply(1:5, function(i) sample(x, 4, prob=p))
     [,1] [,2] [,3] [,4] [,5]
[1,]    4    7    3    8    7
[2,]    2    3    5    4    6
[3,]    6    5    7    5    8
[4,]    7    4    4    9    5
```

- Пакет `samplingbook`

`pps.sampling(z, n, id=1:N, method='sampford', return.PI=F)`

```
> library(samplingbook)
> data <- data.frame(id=1:7, z=c(1.8, 2, 3.2, 2.9, 1.5, 2.0, 2.2))
> set.seed(1)
> pps.sample_sampford <- pps.sampling(z=data$z, n=2, method="sampford", return.PI=F)
> pps.sample_sampford

pps.sampling object: Sample with probabilities proportional to size
Method of Sampford:

PPS sample:
[1] 4 7

Sample probabilities:
            [,1]       [,2]
[1,] 0.37179487 0.06377026
[2,] 0.06377026 0.28205128
```

Аргумент `method` може бити постављен на вредност `'sampford'`, `'tille'`, `'midzuno'`, `'madow'`. Ови методи узорковања разликују се по времену извршења.

Вероватноће укључења првог реда израчунавају се по формули: $\pi_i = n\frac{X_i}{X}$, $i = 1, 2, \ldots, N$.

```
> data(influenza)
> set.seed(90414)
> pps.uzorak <- pps.sampling(z=influenza$population, n=20, method='midzuno')
> pps.uzorak$sample
 [1]  12  20  59 120 166 217 224 226 230 248 267 278 288 313 351 356 369 372 410 411
> (uzorak <- influenza[pps.uzorak$sample,])
       id                        district population cases
12  15082          LK Anhalt-Bitterfeld     184877   121
20   7133              LK Bad Kreuznach     157471    36
59  16061                  LK Eichsfeld     107924     1
120  9675                  LK Kitzingen      89293    55
166 10043                LK Neunkirchen     141426     4
217  8226         LK Rhein-Neckar-Kreis     534989   102
224  3357         LK Rotenburg (Wuemme)     165074    11
226  9277                 LK Rottal-Inn     118800     5
230 16075           LK Saale-Orla-Kreis      90910    71
248  8437               LK Sigmaringen     132419    28
267  7235            LK Trier-Saarburg     141009    47
278  5166                    LK Viersen     303331   104
288  9577   LK Weissenburg-Gunzenhausen      93711     4
313 11010 SK Berlin Marzahn-Hellersdorf     249351    37
351  5513              SK Gelsenkirchen     264765    12
356  2000                    SK Hamburg    1770629   186
369  1002                       SK Kiel     236902    17
372  5114                    SK Krefeld     236516    86
410  9263                  SK Straubing      44625     5
411  8111                  SK Stuttgart     597176   182
> N <- nrow(influenza)
> PI <- pps.uzorak$PI #verovatnoce ukljucenja drugog reda
> htestimate(uzorak$cases, N=N, PI=PI, method='ht')

htestimate object: Estimator for samples with probabilities proportional to size
Method of Horvitz-Thompson:

Mean estimator: 45.55161
Standard Error: 9.218199
```

```
> htestimate(uzorak$cases, N=N, PI=PI, method='yg')

htestimate object: Estimator for samples with probabilities proportional to size
Method of Yates and Grundy:

Mean estimator: 45.55161
Standard Error: 10.49034

> pk <- pps.uzorak$pik[pps.uzorak$sample]
> htestimate(uzorak$cases, N=N, pk=pk, method='hh')

htestimate object: Estimator for samples with probabilities proportional to size
Method of Hansen-Hurwitz (approximate variance):

Mean estimator: 45.55161
Standard Error: 9.760787

> pik <- pps.uzorak$pik #verovatnoce ukljucenja prvog reda
> est.ht <- htestimate(uzorak$cases, N=N, PI=PI, method='ht')
> est.ht$mean*N
[1] 19313.88
> lower <- est.ht$mean*N - qnorm(0.975)*N*est.ht$se
> upper <- est.ht$mean*N + qnorm(0.975)*N*est.ht$se
> #interval poverenja pod pretpostavkom normalne raspodele broja slucajeva prehlade
> c(lower,upper)
[1] 11653.33 26974.44
> #tacan broj slucajeva prehlade
> sum(influenza$cases)
[1] 18900
```

- Систематски узорак са вероватноћом пропорционалном величини (*PPS systematic scheme*)

  На случајан начин свака јединица у популацији означена је по једним природним бројем између 1 и N. Претпоставља се да је за сваку јединицу i позната њена "величина" X_i, $i = 1, 2, \dots, N$. Јединице се систематски бирају у узорак на следећи начин:

  Формирају се кумуланте T_i, $i = 1, 2, \dots, N$; $L = \frac{X}{n}$, а r је случајан број између 1 и N. Јединица означена са i биће укључена у узорак ако $T_{i-1} < r + jL \leq T_i$, за неку вредност $j = 0, 1, \dots, n-1$. Да би се избегла вишеструка понављања исте јединице у узорку пређутно се претпоставља да су јединице за које је $L < X_i$ већ уклоњене из популације и укључене у посебан стратум, који се у потпуности разматра (каже се да он садржи *self-selecting* елементе). Наравно, то изискује да се прерачуна X, након тога израчуна ново L и ново n.

Уколико је потребно овај поступак се понавља све док не престану да се појављују нове *self-selecting* јединице.

Такође, требало би приметити да количник $\frac{X}{n}$ не мора бити цео број, па се тада за L узима најближи цео број. У том случају стварни обим узорка разликоваће се од захтеваног.

На почетку описа ове шеме речено је да су јединице у популацији означене на случајан начин. У пракси, понекад постоји ваљан разлог да јединице у популацији буду уређене у растућем поретку по обележју “величина”.

У пракси помоћно обележје “величина” бира се тако да његова сопствена варијабилност одсликава варијабилност главног обележја y. Тачније, инсистира се да количник Y_i/X_i буде што приближнији константи. На тај начин се постиже мала дисперзија оцене.

- # Пакети pps и sampling

  Функција `inclusionprobabilities(a, n)` (из пакета `sampling`) враћа вероватноће укључења првог реда; `a` је вектор позитивних бројева, `n` је обим узорка.

```
> library(pps)
> library(sampling)
> n <- 2
> z <- data$z
> pik <- inclusionprobabilities(a=z, n=n)
> pik
[1] 0.2307692 0.2564103 0.4102564 0.3717949 0.1923077 0.2564103 0.2820513
```

  Функција `sampfordpi(sizes, n)` (из пакета `pps`) враћа вероватноће укључења другог реда за *Sampford*-ов метод; `sizes` је вектор "величина" јединица у популацији, `n` је обим узорка.

```
> PI_sampford <- sampfordpi(sizes=z, n=n)
> PI_sampford
           [,1]       [,2]       [,3]       [,4]       [,5]       [,6]       [,7]
[1,] 0.23076923 0.03188582 0.05778453 0.05055242 0.02294930 0.03188582 0.03571134
[2,] 0.03188582 0.25641026 0.06516581 0.05704006 0.02594958 0.03602917 0.04033981
[3,] 0.05778453 0.06516581 0.41025641 0.10216684 0.04715868 0.06516581 0.07281474
[4,] 0.05055242 0.05704006 0.10216684 0.37179487 0.04122522 0.05704006 0.06377026
[5,] 0.02294930 0.02594958 0.04715868 0.04122522 0.19230769 0.02594958 0.02907533
[6,] 0.03188582 0.03602917 0.06516581 0.05704006 0.02594958 0.25641026 0.04033981
[7,] 0.03571134 0.04033981 0.07281474 0.06377026 0.02907533 0.04033981 0.28205128
```

  Постоје и одговарајуће функције за друге методе (у пакету `sampling`).

Корисне функције из пакета `pps`:

- `pps1()` – избор једне јединице из популације са *PPS*
- `ppss()` – избор систематског *PPS* узорка
- `ppswr()` – избор *PPS* узорка са понављањем
- ...

```
> example(pps1)

pps1> sizes <- c(9,2,5,17,4,21,15,7,4,11,23,23,14)

pps1> sampleindex <- pps1(sizes)
> sampleindex
[1] 13
> example(ppss)

ppss> sizes <- c(9,2,5,17,4,21,15,7,4,11,23,23,14)

ppss> sampleindices <- ppss(sizes,4)
> sampleindices
[1]  4  7 11 13
> example(ppswr)

ppswr> sizes <- c(9,2,5,17,4,21,15,7,4,11,23,23,14)

ppswr> sampleindices <- ppswr(sizes,4)
> sampleindices
[1]  1  9 12 11
```

Корисне функције из пакета `sampling`:

- `HTestimator()`
- `UPsystematic()` – избор систематског узорка
- `writesample()` – исписује све узорке фиксираног обима

```
> example(HTestimator)

HTstmt> data(belgianmunicipalities)

HTstmt> attach(belgianmunicipalities)
The following objects are masked from belgianmunicipalities (position 3):

    Arrondiss, averageincome, Commune, Diffmen, DiffTOT, Diffwom, INS,
    medianincome, Men03, Men04, Province, TaxableIncome, Tot03, Tot04,
    Totaltaxation, Women03, Women04

HTstmt> # Computes the inclusion probabilities
HTstmt> pik=inclusionprobabilities(Tot04,200)

HTstmt> N=length(pik)

HTstmt> n=sum(pik)

HTstmt> # Defines the variable of interest
HTstmt> y=TaxableIncome

HTstmt> # Draws a Poisson sample of expected size 200
HTstmt> s=UPpoisson(pik)

HTstmt> # Computes the Horvitz-Thompson estimator
HTstmt> HTestimator(y[s==1],pik[s==1])
                [,1]
[1,] 112417271928
```

- Пакет `survey`

  Да би се описао *PPS design* потребно је задати вероватноће узорковања аргументом `prob` функције `svydesign()`, или аргументом `fpc`, али се тада не наводе опционални аргументи `prob`, односно `weight`. Додатно се може показати неопходним да се, аргументом `pps`, одреди *PPS computation* која би требало да буде коришћена (то је уствари апроксимација вероватноћа укључења другог реда). Навођењем опционалног аргумента `variance` може се прецизирати да ли се рачуна *H-T* (`'HT'`) или *Y-G* (`'YG'`) оцена дисперзије. По *default*-у рачуна се *H-T* оцена.

```
> library(survey)
> data(election)
> summary(election$p)
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
0.0000014 0.0007260 0.0022500 0.0086960 0.0057290 0.9037000
> election$votes <- with(election, Bush+Kerry+Nader)
> election$p <- 40*election$votes/sum(election$votes)
> insample <- UPtille(election$p)
> ppsample <- election[insample==1,]
> ppsample$wt <- 1/ppsample$p
> pps.uzorak.design <- svydesign(id=~1, weight=~wt, data=ppsample)
```

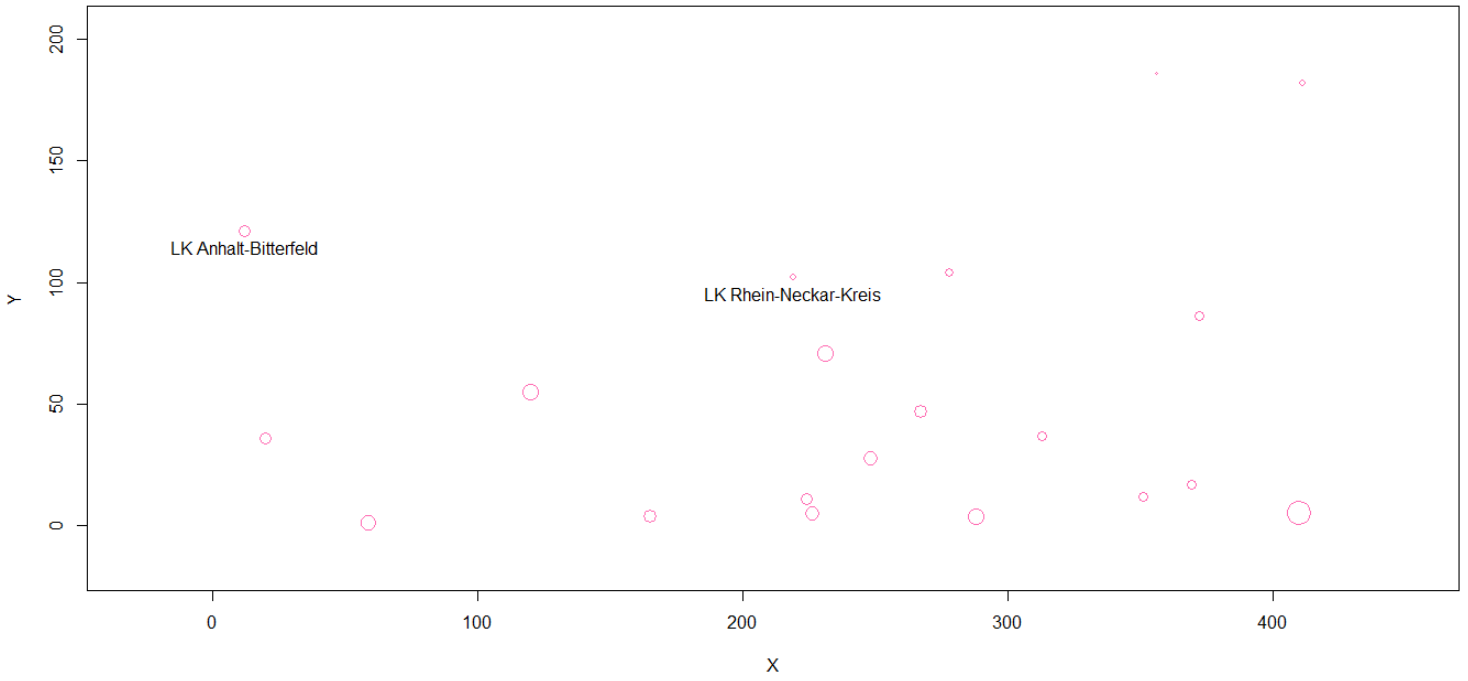
```
> svytotal(~Bush+Kerry+Nader, pps.uzorak.design, deff=T)
          total       SE   DEff
Bush  61204227  2340855 0.0077
Kerry 54421869  2339421 0.0040
Nader   573009    91975 0.0562
```

Аргументом `deff=T` добија се ефекат дизајна (*design effect*), који, на известан начин, показује колики је губитак информације при узорковању. Наиме, *Cluster sampling*, а чак и *PPS* узорковање, имају мању прецизност по опсервацији него узорковање појединачних јединица (нпр. ПСУ).

```
> colSums(election[, 4:6])
    Bush    Kerry    Nader
59645156 56149771   404178
```

Што се графичког приказивања тиче, две су главне потешкоће које се јављају; то су: велики обим узорка и тежине узорковања. Најједноставнији график распршености (*scatter plot*) овде је график "мехурића" (*bubble plot*), уместо тачака исцртавају се кружићи, чије су површине пропорционалне одговарајућим тежинама.

```
> data(influenza)
> str(influenza)
'data.frame':  424 obs. of  4 variables:
 $ id        : int  5354 7131 9771 8425 16077 7132 15081 9171 7331 9371 ...
 $ district  : Factor w/ 424 levels "LK Aachen","LK Ahrweiler",..: 1 2 3 4 5 6 7 8 9 10 ...
 $ population: int  309929 129096 127785 190212 103313 134912 93323 108773 125697 107069 ...
 $ cases     : int  21 14 74 19 36 35 6 39 88 15 ...
> set.seed(90414)
> pps.uzorak <- pps.sampling(z=influenza$population, n=20, method='midzuno')
> ppss.uzorak <- influenza[pps.uzorak$sample,]
> ppss.uzorak$wt <- 1/pps.uzorak$pik[pps.uzorak$sample]
> pps.uzorak.design2 <- svydesign(id=~1, weight=~wt, data=ppss.uzorak)
> svyplot(pps.uzorak.design2$variables$cases ~ pps.uzorak.design2$variables$district,
          design=pps.uzorak.design2, style="bubble", basecol = "hotpink1", inches=.15)
> identify(x=pps.uzorak.design2$variables$district, y=pps.uzorak.design2$variables$cases,
           labels=pps.uzorak.design2$variables$district, n=2)
[1] 1 6
```



- *Sunter's method*

  Овај метод састоји се у томе да се иде редом, по опадајуће уређеном (на основу вредности обележја "величина") списку свих јединица и за свако k (k иде од 1 до N) поступа се на следећи начин:

  - генериши случајан број u_k из $\mathcal{U}(0,1)$ расподеле
  - за $k = 1$, задржи јединицу k у узорку акко је (КОРАК 1)

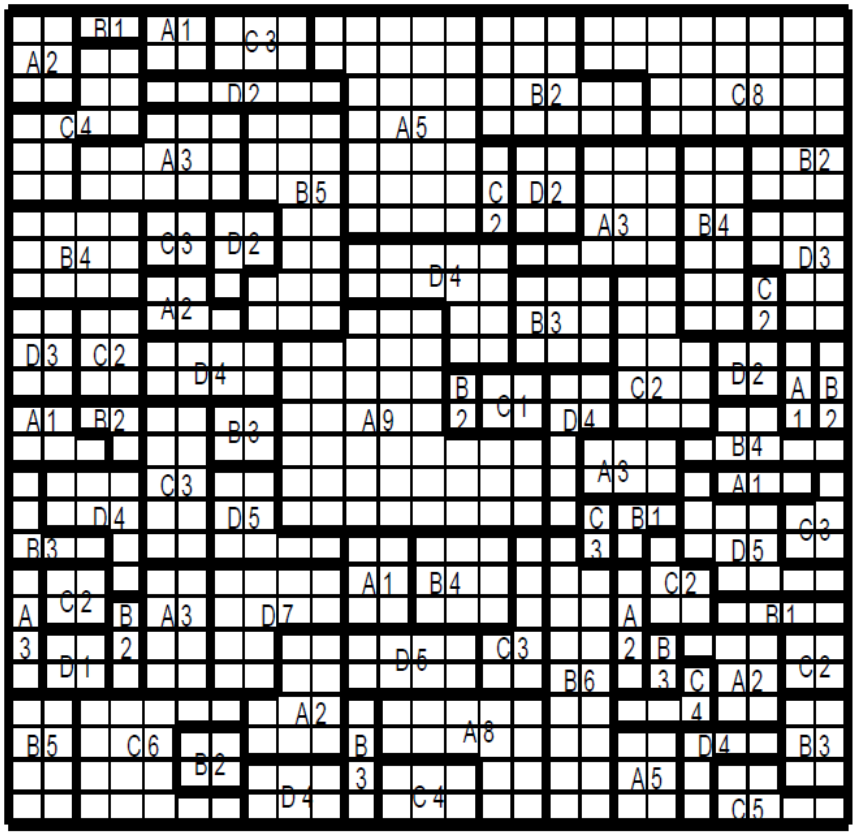
    $u_1 \leq \pi_1$
  - за $k \geq 2$, задржи јединицу k у узорку акко је (КОРАК k)

    $$u_k \leq \frac{n - n_{k-1}}{n - \sum_{i=1}^{k-1} \pi_i} \pi_k$$

    где n_{k-1} представља број јединица већ одабраних у узорак на крају $(k-1)$-ог корака.

- Пример

  Разматра се популација фарми различитих величина и облика, које су смештене на квадратној мрежи димензија 25x25. Свака фарма означена је словом које представља тип фарме и бројем који представља број радника на фарми. Следећа мапа показује границе фарми.



Претпостави да је један квадратић мреже мерна јединица површине. Ако се на случајан начин изабере један квадратић на мрежи, и означи са x_i површина i-те фарме, а са $A = 625$ укупна површина области од интереса, онда је вероватноћа да је изабрана баш i-та фарма једнака

$$p_i = \frac{x_i}{A} = \frac{x_i}{625}$$

Бира се узорак обима $n = 5$ са понављањем и вероватноћама пропорционалним величини. Заправо случајан избор једног пиксела на мрежи постиже се тако што се узоркују по две целобројне вредности (између 1 и 25), које ће представљати координате тог пиксела.

Циљ: оценити укупан број радника; оценити укупан број фарми.

```
> x_koord <- sample(1:25, 5, replace=T)
> y_koord <- sample(1:25, 5, replace=T)
> (uzorak_piksela <- cbind(x_koord, y_koord))
     x_koord y_koord
[1,]      24       19
[2,]      17       19
[3,]       1        3
[4,]      14        8
[5,]      21        6
```

Дакле, одабран је узорак:

Ознака фарме	p_i
$D3$	$10/625$
$D2$	$6/625$
$B5$	$8/625$
$B4$	$9/625$
$A2$	$7/625$

H-H оцена укупног броја радника:

```
> br_radnika <- c(3, 2, 5, 4, 2) #ovo je y obelezje
> p_i <- c(10, 6, 8, 9, 7)/625
> ocena_tot <- sum(br_radnika/p_i)/5
> ocena_tot
[1] 248.5615
```

Стварни подаци:
укупан број радника: 247
укупан број фарми: 78

H-H оцена укупног броја фарми:

```
> ocena_N <- sum(1/p_i)/5 #uzeto je da je obelezje y jednako 1 s.s.
> ocena_N
[1] 80.70437
```