

# Прост случајан узорак (Simple Random Sampling)

3.час

10. март 2016.

# Прост случајан узорак без понављања

## Random Sample Without Replacement - RSWOR

Ово је једна од најједноставнијих и најстаријих метода бирања узорка обима  $n$  из популације која садржи  $N$  јединица.

Прост случајан узорак без понављања је план по коме се  $n$  различитих јединица узорка бира на такав начин да свака могућа комбинација од  $n$  јединица има исту вероватноћу да буде изабрана из популације.

Сваки овакав узорак је комбинација без понављања  $n$ -те класе.

Коришћење таблица случајних цифара или (псеудо)случајних бројева омогућава добијање узорка жељеног обима  $n$ .

## Прост случајан узорак без понављања Random Sample Without Replacement - RSWOR

Нека је  $\Omega$  колекција свих  $2^N$  подскупова од  $S$ , тада је

$$P(s) = \begin{cases} \binom{N}{n}^{-1}, & \text{ако је } n(s) = n \\ 0, & \text{иначе} \end{cases}$$

је план узорковања за прост случајан узорак без понављања.

Овде се сваки од  $\binom{N}{n}$  могућих скупова обима  $n$  са једнаком вероватноћом може одабрати као узорак.

Вероватноће укључења првог и другог реда једнаке су:

$$\pi_i = \frac{n}{N} \quad \pi_{ij} = \frac{n(n-1)}{N(N-1)}$$

# Прост случајан узорак без понављања

## Random Sample Without Replacement - RSWOR

Непристрасна **оцена тотала** обележја популације је

$$\hat{Y}_{RSR} = \frac{N}{n} \sum_{i \in s} Y_i,$$

њена дисперзија је  $V(\hat{Y}_{RSR}) = \frac{N(N-n)}{n} S_y^2$ , где је  $S_y^2 = \frac{1}{N-1} \sum_{i=1}^N (Y_i - \bar{Y})^2$ .

Непристрасна оцена дисперзије  $V(\hat{Y}_{RSR})$  је

$$v(\hat{Y}_{RSR}) = \frac{N(N-n)}{n} s_y^2,$$

где је  $s_y$  узорачки аналогон за  $S_y$ , односно  $s_y^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2$ .

# Прост случајан узорак без понављања

## Random Sample Without Replacement - RSWOR

Непристрасна **оцена средине** обележја популације је

$$\hat{\bar{Y}} = \frac{\hat{Y}_{SRS}}{N},$$

њена дисперзија је  $V(\hat{\bar{Y}}) = \frac{N-n}{Nn} S_y$ .

Непристрасна оцена дисперзије  $V(\hat{\bar{Y}})$  је

$$v(\hat{\bar{Y}}) = \frac{N-n}{Nn} s_y,$$

где је  $s_y$  коригована узорачка дисперзија.

## Прост случајан узорак без понављања

### Random Sample Without Replacement - RSWOR

Количник  $\frac{n}{N}$  је **фракција узорка**, означава се са  $f$  и представља део популације представљен узорком.

Количник  $\frac{N-n}{N}$  је **фактор корекције** због коначности популације.

У пракси, корекција коначне популације се занемарује ако фракција популације не прелази 5%, а у многим случајевима и када је до 10%.

**Популацијска пропорција**  $p$  је удео јединица популације које припадају одређеном нивоу категоричке променљиве од интереса. Са статистичког становишта циљ је оценити параметар  $p$ .

**Индикатор функција** је  $I_i = \begin{cases} 1, & \text{ако } i\text{-та јединица припада нивоу} \\ 0, & \text{иначе} \end{cases}$

# Прост случајан узорак без понављања

## Random Sample Without Replacement - RSWOR

Непристрасна **оцена популацијске пропорције** је

$$\hat{p} = \frac{1}{n} \sum_{i=1}^n I_i,$$

њена дисперзија је  $V(\hat{p}) = \frac{N-n}{Nn} S_y^2$ , где је  $S_y^2 = \frac{N}{N-1} p(1-p)$ .

Непристрасна оцена дисперзије  $V(\hat{p})$  је

$$v(\hat{p}) = \frac{N-n}{N(n-1)} \hat{p}(1-\hat{p})$$

## Прост случајан узорак без понављања Random Sample Without Replacement - RSWOR

$100(1 - \alpha)\%$  интервал поверења за тотал обележја популације је

$$\left[ \hat{Y}_{SRS} - t_{n-1; \frac{\alpha}{2}} \sqrt{\frac{N(N-n)}{n}} s_y^2, \quad \hat{Y}_{SRS} + t_{n-1; \frac{\alpha}{2}} \sqrt{\frac{N(N-n)}{n}} s_y^2 \right]$$

$100(1 - \alpha)\%$  интервал поверења за средину обележја популације је

$$\left[ \hat{\bar{Y}} - t_{n-1; \frac{\alpha}{2}} \sqrt{\frac{N-n}{Nn}} s_y^2, \quad \hat{\bar{Y}} + t_{n-1; \frac{\alpha}{2}} \sqrt{\frac{N-n}{Nn}} s_y^2 \right]$$

Вредност  $t_{n-1; \frac{\alpha}{2}}$  је вредност из таблица за Студентову расподелу са  $n-1$  степени слободе, тако да је  $P\{|t_n| \leq t_{n-1; \frac{\alpha}{2}}\} = 1 - \alpha$ .

Ако је обим узорка већи од 30, вредност  $t_{n-1; \frac{\alpha}{2}}$  чита се из таблица за стандардну нормалну расподелу.



## Прост случајан узорак без понављања

### Random Sample Without Replacement - RSWOR

Код узорковања најчешће се поставља питање који обим узорка би требало одабрати. Одговор није увек једноставан. Генерално, обим узорка би требало планирати тако да се постигне задата тачност оцене.

Величина узорка за оцену средине обележја популације је

$$n = \frac{1}{\frac{1}{n_0} + \frac{1}{N}} = \frac{1}{\frac{d^2}{z^2 s^2} + \frac{1}{N}}$$

Величина узорка за оцену тотала обележја популације је

$$n = \frac{1}{\frac{1}{n_0} + \frac{1}{N}} = \frac{1}{\frac{d^2}{N^2 z^2 s^2} + \frac{1}{N}},$$

где је  $z$  вредност из таблица за стандардну нормалну расподелу, таква да је  $P\{|Z| \leq z\} = 1 - \alpha$ , а  $d$  је највећа апсолутна грешка која се допушта са нивоом поверења  $1 - \alpha$ .

## Прост случајан узорак без понављања Random Sample Without Replacement - RSWOR

Може се приметити да у наведеним формулама за обим узорка фигурише коригована дисперзија  $S_y$ , која најчешће није позната. Оцена тог параметра може се добити нпр. на основу претходног узорка мањег обима. Такође, уколико је обим популације  $N$  велики, може се узети  $n = n_0$ .

Поред овог критеријума одређивања оптималног обима ПСУ задавањем апсолутне грешке оцене, постоје и други критеријуми и то:

- задавањем горње границе дисперзије оцене
- задавањем ширине интервала поверења
- задавањем коефицијента варијације
- задавањем релативне грешке оцене
- задавањем трошкова узорковања.

## Прост случајан узорак са понављањем Random Sample With Replacement - RSWR

Разлика у односу на случајан узорак без понављања јесте у томе што се уместо дефинисања узорка као подскупа популације  $S$ , узорак дефинише као неуређен избор са враћањем јединица из  $S$ .

$$P(s) = \begin{cases} \frac{n}{N^n} \prod_{k \in S} \frac{1}{s_k}, & \text{ако је } n(s) = n \\ 0, & \text{иначе} \end{cases}$$

где је  $s_k$  број понављања  $k$ -те јединице у узорку.

Ако је обим узорка мали у односу на обим популације, значајне разлике између случајног узорка са и без понављања **нема**, јер је вероватноћа понављања било ког елемента мала.

## Прост случајан узорак са понављањем Random Sample With Replacement - RSWR

Вероватноће укључења првог и другог реда једнаке су:

$$\pi_i = 1 - \left(1 - \frac{1}{N}\right)^n \quad \pi_{ij} = 1 - 2 \left(1 - \frac{1}{N}\right)^n + \left(1 - \frac{2}{N}\right)^n$$

Непристрасна **оцена средине** обележја популације је, као и код СУ без понављања, узорачка средина.

Непристрасна оцена дисперзије ове оцене једнака је количнику (кориговане) узорачке дисперзије и обима узорка  $n$ .

Очигледно је да су формуле за дисперзију оцене средине код СУ без и са понављањем приближно исте, када је фракција узорка мала.

# Важни појмови

Стандардна девијација (одступање) је

$$\sqrt{s_y^2}$$

пружа информацију колико су вредности података удаљене (тј. варирају) од своје средње вредности.

Стандардна грешка (средње вредности) узорка обима  $n$  је

$$\sqrt{\frac{s_y^2}{n}}.$$

Стандардна грешка оцене је важан показатељ колико је статистика узорка поуздана оцена параметра.

# Алгоритми за узорковање СУ без понављања

## ① Draw by draw procedure

За свако  $t = 0, 1, \dots, n - 1$  бирамо јединицу  $k$  из популације са вероватноћом

$$q_k = \begin{cases} \frac{1}{N-t}, & \text{ако } k \text{ још увек није изабрана} \\ 0, & \text{иначе} \end{cases}$$

## ② Selection-rejection procedure

За свако  $k = 1, 2, \dots, N$  бирамо јединицу  $k$  из популације са вероватноћом  $\frac{n-j}{N-(k-1)}$  и ако је она одабрана  $j = j + 1$ . Овом процедуром узорак се изабере у једном пролазу кроз податке и представља најбољи алгоритам за избор СУ без понављања.

## ③ Reservoir method

## ④ Random sort procedure

# Алгоритми за узорковање СУ са понављањем

## ① Draw by draw procedure

За свако  $j = 1, 2, \dots, n$  бирамо јединицу  $k$  из популације са вероватноћом  $\frac{1}{N}$ .

## ② Sequential procedure

За свако  $k = 1, 2, \dots, N$  бирамо јединицу  $k$  из популације тачно  $s_k$  пута у складу са биномном расподелом  $\mathcal{B}\left(n - \sum_{i=1}^{k-1} s_i, \frac{1}{N-k+1}\right)$ .

Пример функције за избор СУ обима  $n$  из популације обима  $N$  без понављања:

```
select.reject_srs <- function(N, n){  
  j <- 0  
  pomocni <- runif(N)  
  uzorak <- NULL  
  for(k in 1:N){  
    p <- (n-j)/(N-(k-1))  
    if(pomocni[k] < p){  
      uzorak <- c(uzorak, k)  
      j <- j+1  
    }  
    if(j == n) break  
  }  
  return(uzorak)  
}
```

```
> select.reject_srs(100, 24)
```

```
[1]  8 13 14 17 21 23 28 41 46 48 51 52 54 56 65 66 75 76 80 81 82 87 89 92
```



Генерисање (псеудо) случајног броја из биномне расподеле са задатим параметрима:

```
sl.br.iz_Bin_1 <- function(n, p){  
  sl_br <- 0  
  pomocni <- runif(n)  
  sl_br <- sum(pomocni < p)  
  return(sl_br)  
}
```

```
> rbinom(1, 20, 0.4)
```

```
[1] 9
```

```
> sl.br.iz_Bin_1(20, 0.4)
```

```
[1] 12
```

Пример функције за избор СУ обима  $n$  из популације обима  $N$  са понављањем:

```
sequent_srs <- function(N, n){  
  j <- 0  
  uzorak <- NULL  
  for(k in 1:N){  
    s_k <- sl.br.iz_Bin_1(n-j, 1/(N-k+1))  
    if(s_k > 0)  
      uzorak <- c(uzorak, rep(k, s_k))  
    j <- j+s_k  
    if(j >= n) break  
  }  
  return(uzorak)  
}
```

```
> sequent_srs(100, 24)
```

```
[1]  1 10 19 23 26 31 32 42 44 49 50 53 54 54 63 66 67 74 77 80 82 87 91 99
```

## Функција sample()

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

Ова функција може да пермутује елементе вектора, на случајан начин, не мењајући њихове стварне вредности.

```
> x <- c(2,4,5,6,88,45,65,34,86,8,23,8,43,68,13,19)
> sample(x)
```

```
[1]  4 88 34 65 45  6 86 13  8 68 23  8  5 19  2 43
```

```
> sample(x)
```

```
[1] 68  5 43 23  8 88  6  8 86 19 13 65 34  4  2 45
```

Други и трећи аргумент функције су опциони и њима се може задати величина узорка и да ли елементе узорка бирамо без или са понављањем.

```
> sample(x,5)
```

```
[1] 23  8  6 19  5
```

```
> sample(x,replace = T)
```

```
[1]  4 86  4  6 86 23  8 43  8 68  6 86 68 45  4 23
```

```
> sample(x,replace = T,5)
```

```
[1] 45 45 13  6 19
```

## Пример

У индустријском граду Tamil Nadu у јужној Индији већина породица се бави текстилном индустријом. У следећој табели је дат број машина за ткање у свакој од 51 улице у том месту.

| Редни број | Број машина | Редни број | Број машина | Редни број | Број машина |
|------------|-------------|------------|-------------|------------|-------------|
| 1          | 26          | 18         | 0           | 35         | 31          |
| 2          | 30          | 19         | 27          | 36         | 0           |
| 3          | 0           | 20         | 15          | 37         | 27          |
| 4          | 55          | 21         | 23          | 38         | 11          |
| 5          | 0           | 22         | 0           | 39         | 0           |
| 6          | 0           | 23         | 55          | 40         | 43          |
| 7          | 37          | 24         | 49          | 41         | 55          |
| 8          | 0           | 25         | 0           | 42         | 0           |
| 9          | 0           | 26         | 27          | 43         | 48          |
| 10         | 67          | 27         | 57          | 44         | 29          |
| 11         | 9           | 28         | 54          | 45         | 27          |
| 12         | 35          | 29         | 75          | 46         | 70          |
| 13         | 29          | 30         | 8           | 47         | 13          |
| 14         | 30          | 31         | 45          | 48         | 0           |
| 15         | 0           | 32         | 0           | 49         | 25          |
| 16         | 35          | 33         | 49          | 50         | 27          |
| 17         | 0           | 34         | 0           | 51         | 50          |

Циљ нашег истраживања је одредити укупан број машина за ткање у граду користећи прост случајан узорак величине  $n = 8$ .

## Пример

Користећи стандардне функције из R-а рачунамо следеће:

```
> s <- sort(sample(length(podaci[,1]),8))
> uzorak <- podaci[s,]
> uzorak$Br.Masina
[1] 26 31 11 0 75 0 35 49
> sredina <- mean(uzorak$Br.Masina)
[1] 28.375
> st_devijacija <- sd(uzorak$Br.Masina)
[1] 25.54513
> st_greska <- sd(uzorak$Br.Masina)/sqrt(length(uzorak$Br.Masina))
[1] 9.031567
```

Оцена укупног броја машина и стандардна грешка те оцене добијају се на следећи начин:

```
> Total <- 51
> ocena_Total <- mean(uzorak$Br.Masina)*Total
[1] 1090.125
> disp_ocene <- st_greska*Total
[1] 401.914
```

## Пакет survey

То је стандардни пакет за анализу података добијених истраживањем. Први корак у тој анализи требало би да буде “описивање” плана истраживања (survey design) у R-у.

Функција `svydesign()` узима овај опис и придружује га бази података, како би био креиран survey design object.

Аргументи ове функције су:

- **data**=uzorak - наводимо у ком објекту су смештени подаци
- **id**= 1 - узорковани појединачни објекти, нема кластера
- **fpc**= Total - назначено је која променљива чува обим популације

```
library(survey)
uzorak$Total <- 51
uzorak.design <- svydesign(ids= 1, data=uzorak,fpc= Total)
print(uzorak.design)
```

Independent Sampling design

```
svydesign(ids = ~1, data = uzorak, fpc = ~Total)
```

## Пакет survey

Пакет располаже функцијама за оцењивање средине `svymean()`, тотала обележја `svytotal()` и других параметара. Функцијама се прослеђује име променљиве смештене у survey design object-у. Аутоматски се урачунава фактор корекције због коначности популације.

```
> svymean(~Br.Masina, uzorak.design)
```

|           | mean  | SE     |
|-----------|-------|--------|
| Br.Masina | 26.25 | 4.6067 |

```
> svytotal(~Br.Masina, uzorak.design)
```

|           | total  | SE     |
|-----------|--------|--------|
| Br.Masina | 1090.1 | 369.05 |

Интервал поверења се може оценити помоћу функције `confint()`.

```
> confint(svytotal(~ Br.Masina, uzorak.design))
```

|           | 2.5 %    | 97.5 %   |
|-----------|----------|----------|
| Br.Masina | 878.2751 | 1799.225 |

## Пакет survey

Ако се не назначи обим популације, потребно је задати вероватноће узорковања (sampling probabilities) или тежине узорковања (sampling weights).

```
> pw <- 51/8  
> uzorak$pw <- pw  
> uzorak.design.nofpc <- svydesign(ids= 1, data=uzorak, fpc= Total, weights= pw)  
> uzorak.design.nofpc
```

Independent Sampling design

```
svydesign(ids = ~1, data = uzorak, fpc = ~Total, weights = ~pw)
```

```
> svytotal( Br.Masina, uzorak.design.nofpc)
```

|           | total  | SE     |
|-----------|--------|--------|
| Br.Masina | 1090.1 | 369.05 |



У презентацији коришћени:

- Љ. Петровић, *Теорија узорака и планирање експеримената*, Економски факултет, Универзитет у Београду, 2007
- S. Sampath, *Sampling Theory and Methods*, Alpha Science International Ltd., Harrow, U.K. 2005
- [http://www.ats.ucla.edu/stat/faqsvy\\_r\\_scpsu.htm](http://www.ats.ucla.edu/stat/faqsvy_r_scpsu.htm)