ТЕОРИЈА УЗОРАКА 2

12. 04. ’13.

ВЕЖБАЊА

- Написати функције за бирање елемената популације обима N у узорак обима n, код простог случајног узорка, користећи алгоритме:
  - Draw by draw procedure for SRS/SRSWOR

Definition j : Integer;

For $t = 0, 1, \ldots, n-1$ do

одабери јединицу k из популације, са вероватноћом

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

  - Selection-rejection procedure for SRS/SRSWOR

Definition k, j : Integer;

$j = 0$

For $k = 1, 2, \ldots, N$ do

са вероватноћом $\frac{n-j}{N-(k-1)}$ одабери јединицу k; $j = j + 1$

- Написати функције за бирање елемената популације обима N у узорак обима n, код случајног узорка са понављањем, користећи алгоритме:
  - Draw by draw procedure for SRSWR

Definition j : Integer;

For $j = 1, 2, \dots, N$ do

одабери јединицу из популације, са (једнаком) вероватноћом $\frac{1}{N}$

  - Sequential procedure for SRSWR

Definition k, j : Integer;

$j = 0$

For $k = 1, 2, \dots, N$ do

одабери јединицу k тачно s_k пута у складу са биномном расподелом

$$\mathcal{B}\left(n - \sum_{i=1}^{k-1} s_i, \frac{1}{N-k+1}\right)$$

- Узет је прост случајан узорак од 10 кућа из популације од 100 кућа. Број становника у кућама из узорка је: 2, 5, 1, 4, 4, 3, 2, 5, 2, 3.

  а) Оценити просечан број становника по кући и одредити оцену варијансе те оцене.

  б) Оценити укупан број становника у популацији и одредити оцену варијансе те оцене.

  в) Одредити 90% интервал поверења за популацијске вредности средине и тотала обележја.

- Ботаничар жели да оцени број дрвета брезе у некој области. Област је подељена на 1000 подобласти. Из претходних испитивања познато је да је дисперзија у броју стабала по области приближно 45. Код простог случајног узорка, одредити величину узорка тако да 95% интервал поверења за оцену укупног броја дрвета у области, која се проучава не буде већи од:

  а)500 дрвета б)1000 дрвета в)2000 дрвета

ОЦЕЊИВАЊЕ ПРОПОРЦИЈЕ (ATTRIBUTE PROPORTION ESTIMATION)

- Претпостави се да је истраживач у току истраживања заинтересован за неки атрибут (карактеристику), везан за јединице популације.
- **Популацијска пропорција** p (Population proportion) је пропорција (удео) јединица популације који имају тај, поменути атрибут.
- Са статистичког становишта циљ је оценити параметар p.
- Индикатор функција:

$$y_i = \begin{cases} 1, & \text{ако јединица } i \text{ поседује атрибут} \\ 0, & \text{иначе} \end{cases}$$

- Оцена пропорције:

  Код простог случајног узорка обима n је $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} y_i$

  непристрасна оцена популацијске пропорције, а њена дисперзија је

$$V[\hat{p}] = \frac{N-n}{Nn} S_y^2$$

  где је $S_y^2 = \frac{N}{N-1} \cdot p(1-p)$

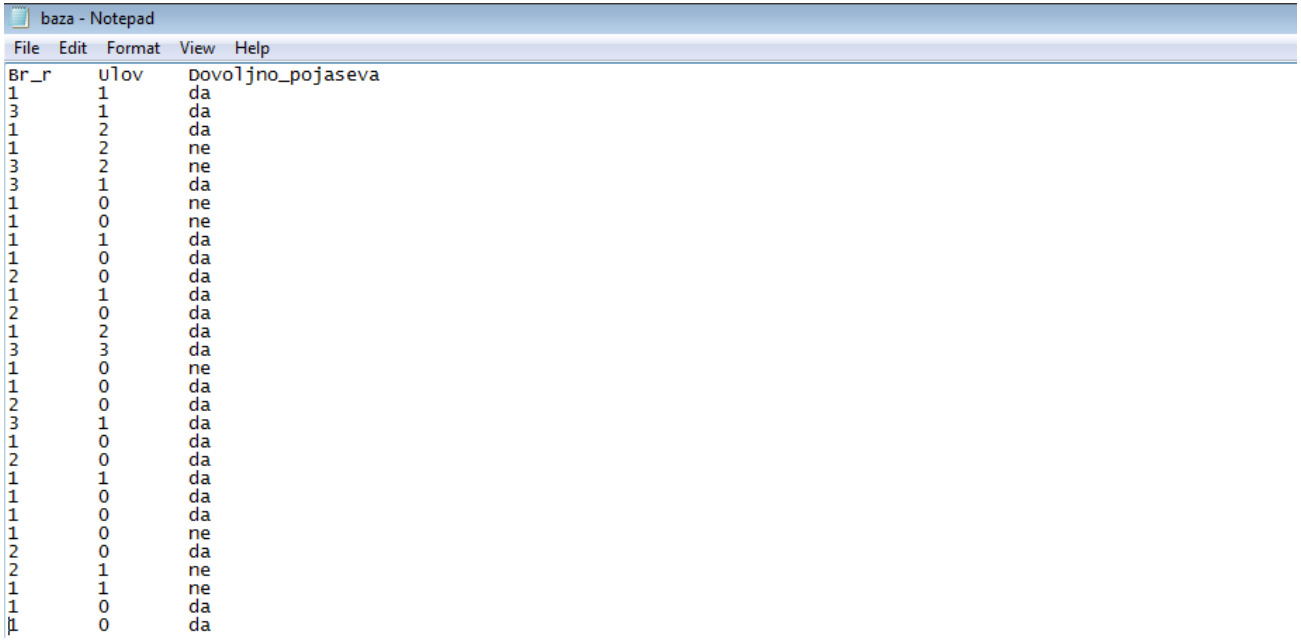
- Непристрасна оцена $V[\hat{p}]$ је $v[\hat{p}] = \frac{N-n}{N(n-1)} \cdot \hat{p}(1-\hat{p})$

- Квадратни корен од $V[\hat{p}]$ је **стандардно одступање** (standard deviation) оцене $\hat{p}$.
- Квадратни корен од $v[\hat{p}]$ је **стандардна грешка** (standard error) оцене $\hat{p}$.

ДОДАТНО ВЕЖБАЊЕ

- Рекреативни риболов

  Дата је следећа база података - узорак обима 30 (популација је обима 168): свака врста у бази одговара регистрованим подацима за по један чамац.


baza - Notepad

File Edit Format View Help

Br_r	Ulov	Dovoljno_pojaseva
1	1	da
3	1	da
1	2	da
1	2	ne
3	2	ne
3	1	da
1	0	ne
1	0	ne
1	1	da
1	0	da
2	0	da
1	1	da
2	0	da
1	2	da
3	3	da
1	0	ne
1	0	da
2	0	da
3	1	da
1	0	da
2	0	da
1	1	da
1	0	da
1	0	da
1	0	ne
2	0	da
2	1	ne
1	1	ne
1	0	da
1	0	da

- Циљ: оценити укупан број риболоваца који су учествовали у рекреативном пецању; оценити просечан број риболоваца на сваком чамцу; оценити популацијску пропорцију бродова који су имали довољно појасева за спасавање за посаду на броду; одредити оцене варијанси оцена и стандардне грешке оцена; одредити 95% интервале поверења.
- t.test()
- prop.test()

У R-у – пакет SURVEY

RGui - [R Console]

File Edit View Misc Packages Windows Help

```
> library(survey)
> baza<-read.table("C:/Users/Lenchy/Desktop/baza.txt",header=T)
> baza$N <- 168 #dodata je kolona bazi koja sadrzi obim populacije
>
> #f-jom 'svydesign()' se definise plan istrazivanja (Survey Design)
>
> primer.design <- svydesign(data=baza, ids=~1, #nema klastera, tj. smatra se da je u pitanju SRS
+ variables=~Br_r+Ulov+Dovoljno_pojaseva, #znakom ~ se identifikuju promenljive u bazi koje treba analizirati
+ fpc=~N) #ukupan broj camaca
> print(primer.design)
Independent Sampling design
svydesign(data = baza, ids = ~1, variables = ~Br_r + Ulov + Dovoljno_pojaseva,
    fpc = ~N)
>
> #f-jama 'svymean()', 'svytotal()' se mogu ocenjivati, redom, sredina i total obelezja populacije
> #ove f-je obicno zahtevaju promenljive koje treba analizirati, bazu koja sadrzi podatke i objekat koji "cuva" plan istrazivanja, a prethodno je kreiran
> (o.sredina <- svymean(~Br_r+Ulov+Dovoljno_pojaseva, primer.design))
                        mean     SE
Br_r                 1.53333 0.1284
Ulov                 0.66667 0.1397
Dovoljno_pojasevada 0.73333 0.0744
Dovoljno_pojasevane 0.26667 0.0744
> (o.sredina.ip <- confint(o.sredina))
                          2.5 %    97.5 %
Br_r                 1.2816361 1.7850305
Ulov                 0.3928823 0.9404510
Dovoljno_pojasevada 0.5874623 0.8792044
Dovoljno_pojasevane 0.1207956 0.4125377
> (o.total <- svytotal(~Br_r+Ulov, primer.design))
     total     SE
Br_r 257.6 21.574
Ulov 112.0 23.468
> (o.total.ip <- confint(o.total))
         2.5 %   97.5 %
Br_r 215.31487 299.8851
Ulov  66.00423 157.9958
>
```

УЗОРАК СА НЕЈЕДНАКИМ ВЕРОВАТНОЋАМА (UNEQUAL PROBABILITY SAMPLING)

- Узорак са неједнаким вероватноћама се често 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.

- За такве популације, 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 једнако 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 одражава 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: $T_1 = X_1, T_2 = X_1 + X_2, \ldots, T_N = X_1 + X_2 + \cdots + X_N$

  Потом се бира 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 за популације великог обима.

- Lahiri's Method

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

  КОРАК 1 Одабери случајан број i између 1 и N

  КОРАК 2 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 1 и 2 док се не одабере јединица.

Hansen-Hurwitz-ова оцена

- 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,

  $i = 1, 2, \ldots, n$.

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

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

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

- 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
- Оцена средине:

Ако је $\hat{\bar{Y}} = \frac{\hat{Y}_{H-H}}{N}$, онда је, код овог узорка, $\hat{\bar{Y}}$ непристрасна оцена средине обележја популације, а непристрасна оцена њене дисперзије је

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

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 (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

- 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:

  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

$$v\left[\hat{Y}_{H-T}\right] = \sum_{i=1}^{N}\sum_{j=1}^{N} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j \pi_{ij}} Y_i Y_j I_i I_j$$

  Ако се са s' означи редуковани узорак (из почетног узорка се избаце јединице које се понављају), еквилалентне формуле за оцене су

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

- Требало би приметити да претходни израз за оцену дисперзије може бити негативан. Постоји алтернативна оцена, која је увек ненегативна.
- Код PPSWOR:

  Може се користити Horvitz-Thompson-ова оцена, под претпоставком да су вероватноће укључења познате. Код овог узорка нису доступни експлицитни изрази за вероватноће укључења. Уз помоћ рачунара, истраживач може да направи списак свих могућих резултата извлачења n јединица у узорак и да срачуна вероватноће укључења.
- Код PPSWR:

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

У R-у

- пакет под називом pps – садржи функције за одабир узорака са вероватноћама пропорционалним величини итд.
- sample() - опционални аргумент prob=

The optional `prob` argument can be used to give a vector of weights for obtaining the elements of the vector being sampled. They need not sum to one, but they should be non-negative and not all zero. If `replace` is true, Walker's alias method (Ripley, 1987) is used when there are more than 250 reasonably probable values: this gives results incompatible with those from R < 2.2.0, and there will be a warning the first time this happens in a session.

If `replace` is false, these probabilities are applied sequentially, that is the probability of choosing the next item is proportional to the weights amongst the remaining items. The number of nonzero weights must be at least `size` in this case.

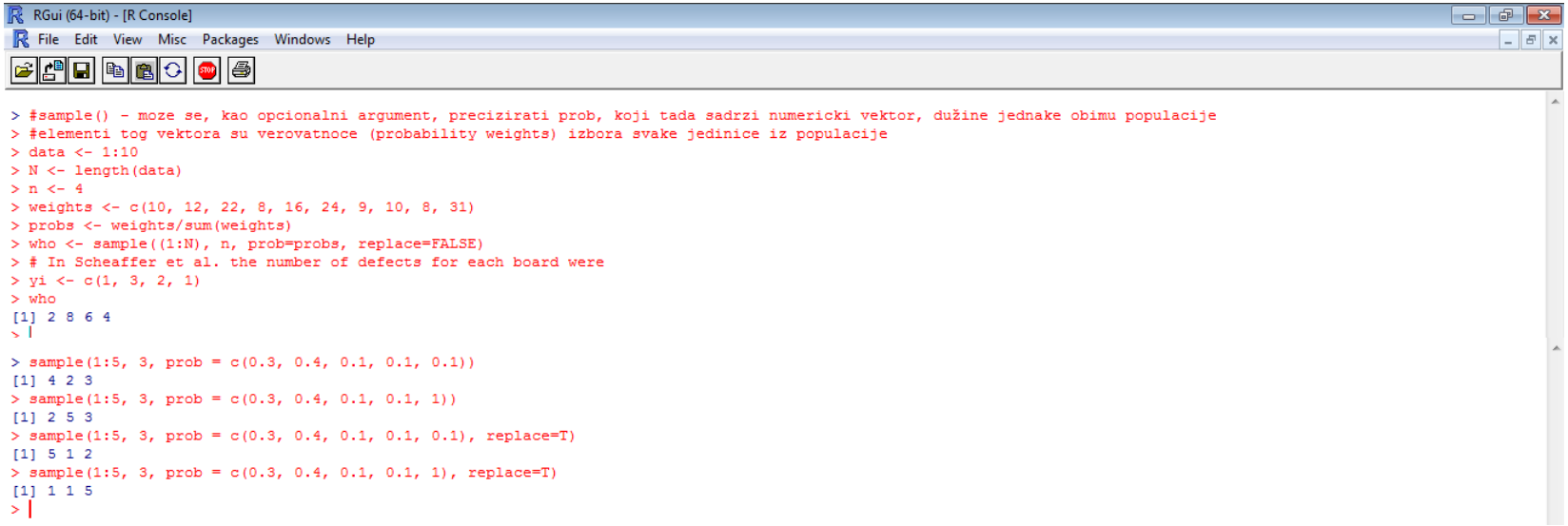
- Пример:

Example (from *Scheaffer* et al., p. 80):

An investigator wishes to estimate the average number of defects per keyboard on keyboards of electronic components manufactured for installation in computers. The keyboards contain varying numbers of components, and the investigator feels that the number of defects should be positively correlated with the number of components on a keyboard.

Thus, *pps* sampling is used with the probability of selecting anyone keyboard for the sample being proportional to the number of components on that keyboard. A sample of $n = 4$ keyboards is to be selected from the $N = 10$ keyboards of one day's production. The number of components on the 10 keyboards are, respectively: 10, 12, 22, 8, 16, 24, 9, 10, 8, 31.

After the sampling was completed, the number of defects found on the four keyboards were, respectively, 1, 3, 2 and 1. Estimate the average number of defects per keyboard, and place a bound on the error of estimation.


RGui (64-bit) - [R Console]
File Edit View Misc Packages Windows Help
> #sample() - moze se, kao opcionalni argument, precizirati prob, koji tada sadrzi numericki vektor, dužine jednake obimu populacije
> #elementi tog vektora su verovatnoce (probability weights) izbora svake jedinice iz populacije
> data <- 1:10
> N <- length(data)
> n <- 4
> weights <- c(10, 12, 22, 8, 16, 24, 9, 10, 8, 31)
> probs <- weights/sum(weights)
> who <- sample((1:N), n, prob=probs, replace=FALSE)
> # In Scheaffer et al. the number of defects for each board were
> yi <- c(1, 3, 2, 1)
> who
[1] 2 8 6 4
> sample(1:5, 3, prob = c(0.3, 0.4, 0.1, 0.1, 0.1))
[1] 4 2 3
> sample(1:5, 3, prob = c(0.3, 0.4, 0.1, 0.1, 1))
[1] 2 5 3
> sample(1:5, 3, prob = c(0.3, 0.4, 0.1, 0.1, 0.1), replace=T)
[1] 5 1 2
> sample(1:5, 3, prob = c(0.3, 0.4, 0.1, 0.1, 1), replace=T)
[1] 1 1 5
>