Данијел Алексић
Неки делови дословно преузети из материјала Нађе Обреновић

Увод у теорију узорака

Вежбе које прате предавања доц. др Ленке Главаш

Универзитет у Београду
Математички факултет

Садржај

Предговор

Ова скрипта настала је[1] као резултат мог држања вежби из Увода у теорију узорака на Математичком факултету Универзитета у Београду. Није замишљено да ови материјали буду свеобухватни за основе Теорије узорака, већ представљају само практични део који прати теоријску основу са предавања. То значи да ће неки појмови да буду подразумевани, а њихова разјашњења могу се наћу у литератури за предавања. Ознаке су формиране тако да прате предавања доц. др Ленке Главаш, која је предметни професор на курсу у тренутку писања ових редова.

Основа скрипте јесу материјали колегинице Нађе Обреновић, која је овај курс држала годину дана пре мене. Овом приликом јој се захваљујем. Хвала и колегиници Мирјани Вељовић, на основу чијих материјала су настали Нађини материјали. Захваљујем се и предметној професорки на корисним сугестијама, као и доц. др Милану Јовановићу код кога сам полагао Теорију узорака.

Молио бих сваког ко након мене буде држао овај курс да ми пише на мејл `danijel.aleksic98@gmail.com` како бих му послао изворни код скрипте да је додатно унапређује и, наравно, допише себе као аутора. Ако у тренутку читања ових редова и даље држим овај курс, молим да ми се за све грешке у скрипти обрати на `danijel_aleksic@matf.bg.ac.rs`.

У Београду,
Љета Господњег 2022,
Сочинитељ.

[1] Одбијам да прихватим да је *скрипта* реч у множини.

Схема боја

Теорема, лема, став

Обрати пажњу

Напомена

Пример

Дефиниција

Задатак

Вежбе 1

Теорија узорака: основни појмови

1.1 Неопходно предзнање

Напомена. Није неопходно, али је више него препоручљиво, бити упознат са првим предавањем пре читања даљег текста.

Пример 1.1 (измишљен!(стварно))**.** Прича се да је групи психолога за истраживање био неопходан просечан коефицијент интелигенције студената на Математичком факултету у Београду. У ту сврху на поменутом факултету платили су сараднику у настави са 50% радног времена[a] да сваком студенту МАТФа да тест интелигенције и достави добијене резултате. Круже приче да се десила једна од три ствари:

1. Поменути сарадник заиста је одрадио свој посао и доставио неопходне податке.
2. Мрзело га је да анкетира баш сваког студента, већ је на случајан начин одабрао 100, њима дао тест и измерио IQ, те њихов просек представио као просек на целом факултету, у нади да неће бити „провађен".
3. И тај посао био му је превише. Седео је на ходнику, и првих 100 студената који су наишли добили су IQ тест. Многи су наишли више пута.

У овој ситуацији скуп свих студената МАТФа звали бисмо **популација**. Коефицијент интелигенције звали бисмо **обележје** на популацији, а његов просек на целој популацији јесте пример једног **параметра популације**. У случају 1. кажемо да је сарадник извршио **попис** (енг. *census*), јер је измерио вредност обележја на свакој јединки из популације.
У случајевима 2. и 3. вредност обележја није мерена на баш свим јединкама, већ само на неким. Тада кажемо да је вршено извлачење узорка, тј. **узорковање** (енг. *sampling*). У случају 2. узорковало се без понављања (све јединке у узорку су различите), а у случају 3. са понављањем.

[a] Свака сличност са реалним ликовима је случајна.

Сада ћемо мало формализовати горе речено. Нека имамо неки скуп јединки које посматрамо, а који ћемо означити са $\Omega = \{\omega_1, \dots, \omega_N\}$. Овај скуп зове се **популација**. У претходном примеру видели смо да популација не мора бити скуп бројева, већ је то просто неки скуп, код нас коначан. Ми на популацији желимо да меримо вредност неког обележја, те **обележје** формално дефинишемо као неку функцију $Y : \Omega \to \mathbb{R}$, која свакој јединки популације додељује неки реалан број - вредност обележја на тој јединки. Отуда добијамо скуп $\{y_1, \dots, y_N\}$ који представља скуп вредности обележја на свим јединкама.

Претпостављамо да су све јединке поређане у коначан низ, то јест да се зна која је прва, друга, трећа итд. јединка. Тада је y_k вредност обележја на k-тој јединки. Више о томе како и зашто можемо то да претпоставимо - на предавањима.

У претходном примеру мерили смо просек обележја на популацији. То је само један од могућих параметара на популацији. Уопште, **параметар** популације јесте свака функција $\mathbb{R} \ni \theta = f(y_1, \dots, y_n)$.

Параметара популације у начелу има бесконачно много, али ово су неки од најкоришћенијих:

- **Популацијска средина**:
$$m_Y = m = \frac{1}{N}\sum_{k=1}^{N} y_k.$$
- **Популацијски тотал:**
$$\tau_Y = \tau = \sum_{k=1}^{N} y_k.$$
- **Популацијска дисперзија:**
$$\sigma_Y^2 = \sigma^2 = \frac{1}{N-1}\sum_{k=1}^{N}(y_k - m_Y)^2.$$

Обрати пажњу: Зашто се дели са $N-1$, када је смислено да се дели са N?

- **Стандардно одступање:**
$$\sigma_Y = \sigma = \sqrt{\sigma_Y^2}.$$

Откуд потреба за стандардним одступањем, када је оно једнозначно одређено дисперзијом? Хинт: Ако обележје меримо у метрима, у којим јединицама је дисперзија, а у којим стандардно одступање?

Из многих разлога, који су детаљније обрађени на предавањима, веома често није могуће израчунати вредност параметра на свим јединкама популације. Стога се обично бира неки подскуп популације, те се на њему рачуна параметар. Тај процес одабира зове се **узорковање**. Ми ћемо се овде бавити **вероватносним** узорковањем, што значи да сваки узорак има неку вероватноћу да буде изабран, а узорковање се врши у складу са расподелом вероватноћа која је дефинисана на скупу свих могућих узорака на популацији (обавезно видети предавања!!!).

Означимо са S скуп индекса оних јединки које су „упале'' у узорак, а са $n(S)$ број елемената тог узорка. Како нам је у интересу да што боље проценимо параметар на основу узорка, ми ћемо за то користити **оцену**, која заправо није ништа друго до нека функција која за аргументе има само елементе узорка. Ако смо параметар означили са θ, пракса је да му оцену означимо са $\hat{\theta}$.

Статистика је свака[a] реална функција узорка. Статистика која оцењује параметар зове се оцена. Користићемо оба термина без опасности од забуне.

[a] Иначе мора бити тзв. Борелова, али за коначну популацију је небитно.

Горе смо навели најкоришћеније параметре, а сада ћемо навести њихове најчешће коришћене оцене:

- **Узорачка средња вредност:**

$$\overline{Y} = \frac{1}{n(S)} \sum_{k \in S} y_k.$$

- **Узорачки тотал:**

$$T = n(S)\overline{Y}.$$

- **Узорачка дисперзија:**

$$\overline{S}^2 = \frac{1}{n(S)-1} \sum_{k \in S} (y_k - \overline{Y})^2.$$

Овде је место за још једну напомену. Могло би се погрешно закључити да је вероватносни модел у коме радимо онај у коме је простор исхода једнак скупу јединки, а да је статистика функција која слика скуп јединки у реалне бројеве. **То није**.
Нема ништа случајно у вредности обележја на било којој од јединки. Оно је фиксан број, макар ми и не знали колико је. Случајан је **избор узорка**. Дакле, простор исхода је заправо простор свих узорака на популацији, а оцена је функција која узме узорак и врати реалан број. Ово је малкице другачији приступ од оног са курса Статистика, па се саветује опрез (Заправо није другачији, али је довољно специфичан да заслужује напомену).
Рецимо, ако је на k-том узорку вредност оцене θ_k, а вероватноћа избора тог узорка једнака p_k, очекивање оцене је $\sum p_k \theta_k$. О вероватноћама избора биће више речи у наставку курса.

Наравно, ми средњу вредност можемо оценити нулом, увек. И то би се звало оцена, али јако бесмислена. Ми желимо да наше оцене буду смислене, те им на неки начин морамо мерити квалитет.

Оцена $\hat{\theta}$ параметра θ је **непристрасна** уколико је $\mathbf{E}\hat{\theta} = \theta$. **Пристрасност** оцене дефинише се као $B(\hat{\theta}) = \mathbf{E}\hat{\theta} - \theta$.
Средњеквадратна грешка оцене дефинише се као $\mathbf{MSE}\hat{\theta} = \mathbf{E}(\hat{\theta} - \theta)^2$. Када поредимо две оцене, **боља** је она која има мању средњеквадратну грешку.

Средњеквадратна грешка и дисперзија у општем случају нису исте. За једну класу оцена, пак, јесу. За коју?

1.1.1 Прост случајан узорак (ПСУ)

Узорковање (извлачење узорка) се може вршити на много начина. У зависности од тога на који начин се врши, разликујемо различите типове узорковања. Један од основних јесте и **прост случајан узорак**. Реч прост потиче од чињенице да је вероватноћа избора сваког од узорака иста. У зависности од тога да ли у узорку може бити понављања (јединки, не вредности обележја!) разликујемо два типа ПСУ:

1. **ПСУ без понављања.** Нека је n унапред задат обим узорка. Под узорком подразумева се сваки подскуп популације који има n елемената. Таквих узорака има $\binom{N}{n}$, а вероватноћа избора сваког од њих једнака је

$$\frac{1}{\binom{N}{n}}.$$

Треба извући 5 карата из шпила. Извучемо прву, забележимо вредност и знак, и **не вратимо** је у шпил. Из остатка шпила извучемо другу; не вратимо ни њу. Радњу поновимо 5 пута. Међу 5 извучених карата нема истих.

2. **ПСУ са понављањем.** Нека је n и овде унапред дат обим узорка. Под узорком овде ћемо посматрати сваку n-варијацију са понављањем елемената популације. Овде узорак није скуп, већ уређена n-торка.

   Истина је да смо негде раније рекли да је узорак сваки подскуп популације, што би имплицирало да је то скуп, а овде видимо да он може да буде и уређена n-торка. Нећемо се оптерећивати тим формалностима у овом тренутку. Овде ћемо ићи на поверење: верујте ми да може да се формализује.

   ПСУ са понављањем је једини тип узорка са којим смо се сусрели на курсу Статистика на трећој години. На курсу Вероватноћа, у почетном делу курса, сусрели смо се са разним типовима узорака, али нисмо знали да их препознамо, већ смо ишли на осећај. Сада ћемо научити и то.

   Треба извући 5 карата из шпила. Извучемо прву, забележимо вредност и знак, и **вратимо** је у шпил. Вучемо другу, вратимо и њу. Радњу поновимо 5 пута. Међу 5 извучених карата може бити истих.

   Таквих узорака има N^n, а вероватноћа избора сваког од њих једнака је

$$\frac{1}{N^n}.$$

Уколико вероватноће избора свих узорака нису једнаке, кажемо да је узорак случајан, али није прост. Тада се ствари незнатно мењају.

1.2 Задаци

Задатак 1.1. Колико има простих случајних узорака без понављања, а колико са понављањем, ако се вади узорак обима 7 из популације која садржи 20 јединки?

Решење. Овај задатак је рутинског типа и за загревање. Наравно, знамо ручно израчунати биномни коефицијент, те то нема потребе да спроводимо. Решићемо задатак у R-у, да се подсетимо тог језика (или упознамо с њим).

Прво ћемо уписати обим популације и обим узорка.

```
N <- 20
n <- 7
```

Простих случајних узорака без понављања има $\binom{N}{n} = \binom{20}{7}$, што се може добити командом

```
choose(N, n)
```

```
## [1] 77520
```

Слично добијамо и колико има ПСУ са понављањем

```
N^n
```

```
## [1] 1.28e+09
```

Добили смо резултат у експоненцијалним ознакама. Можемо се тога решити:

```
options(scipen=999) # uklanja e^ notaciju (osnova je 10, nije e)
N^n
```

```
## [1] 1280000000
```

Ако хоћемо старе ознаке назад, просто ћемо позвати `options(scipen=0)`.

Задатак 1.2. Марко има 7 јабука, Петар 2, Јован 2 и Саша 6. Узорак обима 2 бира се тако да је вероватноћа да су Марко и Петар у узорку једнака 1/3, Марко и Јован 1/2, Петар и Јован 1/6, док је вероватноћа избора свих осталих узорака једнака нули. Као оцена укупног броја јабука користи се статистика $\hat{t} = N\overline{Y}$. Да ли је ова оцена непристрасна?

Прво ћемо задатак решити класично (ручно), а затим у R-у.

Решење - ручно. Задатак ћемо решити у три корака: (1) срачунаћемо праву вредност суме броја јабука; (2) наћи ћемо расподелу вероватноћа оцене и (3) на основу те расподеле срачунати очекивање оцене. Ако буде једнако правој вредности, оцена ће бити непристрасна, иначе не. Кренимо

(1) Уочимо да је код нас $N = 4$ и $n = 2$. Одлучимо се да јединке популације нумеришемо као (Марко, Петар, Јован, Саша) $= (1, 2, 3, 4)$. Тада је $y_1 = 7, y_2 = 2, y_3 = 2$ и $y_4 = 6$, па је $t = y_1+y_2+y_3+y_4 = 17$ (користимо t уместо τ да нам се ознаке слажу са онима у програму касније).

(2) Расподела вероватноћа оцене дата је у доњој табели:

узорак	{М, П}	{М, Ј}	{П, Ј}	остали
$\hat{t}$	$4 \cdot \frac{7+2}{2} = 18$	18	8	$*$
вероватноћа	1/3	1/2	1/6	0

Није нам битно колико је $*$ на било ком од осталих узорака, јер нама треба очекивање, па ћемо свакако множити нулом тај број.

(3) Сада рачунамо очекивање.

$$\mathbf{E}\hat{t} = \frac{18}{3} + \frac{18}{2} + \frac{8}{6} + 0 \cdot * = 6 + 9 + \frac{4}{3} = \frac{49}{3} \approx 16.3333\ldots .$$

Ово није једнако правој вредности 17, па оцена није непристрасна.

Решење - програм. Сада ћемо видети како се горње решење имплементира у R-у.

```
# Upisujemo populaciju (Marko, Petar, Jovan, Saša)
pop <- c(1, 2, 3, 4)
# Obim populacije
N <- length(pop)
# Vrednosti obeležja na svim jedinkama
Y <- c(7, 2, 2, 6)
# U listu upisujemo samo one uzorke koji imaju nenula verovatnoće izbora
uzorci <- list(pop[c(1, 2)], pop[c(1, 3)], pop[c(2, 3)])
# Verovatnoće izbora
P_uzorka <- c(1/3, 1/2, 1/6)

# U ovaj vektor upisujemo uzoračke srednje vrednosti,
```

```r
# jer su nam ocene oblika t_ocena = N * Y_sr
Y_sr <- c() # on je sada prazan
# Sada ćemo proći kroz svaki uzorak i sračunati srednju vrednost na njemu
for (i in 1:length(uzorci)) {
  Y_sr[i] <- mean(Y[uzorci[[i]]]) # Elementu liste se pristupa
                                  # dvama uglastim zagradama
}
# Ocenu dobijamo kada svaku sredinu pomnožimo veličinom populacije
t_ocena <- N * Y_sr
# Sada računamo očekivanje
E_t_ocena <- sum(t_ocena * P_uzorka)
E_t_ocena
```

```
## [1] 16.33333
```

```r
# Možemo proveriti da li je nepristrasna
E_t_ocena == sum(Y)
```

```
## [1] FALSE
```

Ocena је пристрасна.

Задатак 1.3. Испитати да ли је боља (у средњеквадратном смислу) оцена $\hat{t} = N\overline{Y}$ укупне суме обележја на основу узорка обима 2, или на основу узорка обима 3, из популације $\{1, 2, 3, 4\}$, ако се користи прост случајан узорак без понављања. Обележје популације је редни број јединке.

Решење. Од сада ћемо задатке решавати само у R-у. Јасно, треба срачунати средњеквадратне грешке за обе оцене и упоредити их.

```r
# Populacija
pop <- c(1, 2, 3, 4)
# Obeležje je redni broj jedinke
Y <- pop
# Obim populacije
N <- length(pop)
# Veličine uzorka u oba slučaja
n1 <- 2
n2 <- 3
# Funkcija combn(vektor, dužina) pravi sve kombinacije bez ponavljanja
# elemenata iz "vektor", dužine "dužina"
M1 <- combn(pop, n1) # Svaka kolona je jedna kombinacija
# Sada upisujemo sve uzorke obima n1
psu_1 <- list()
for (i in 1:choose(N, n1)) {
  psu_1[[i]] <- M1[, i] # i-ta kolona
}
# Naravno, radimo dupli posao jer su svi uzorci već u matrici, ali ovako
# je didaktički.

# Sve verovatnoće izvlačenja su iste jer imamo PSU, ali ćemo ih upisati u vektor
# radi lakšeg računanja očekivanja
P_uzorka_1 <- rep(1/choose(N, n1), choose(N, n1))
```

```r
# Sve isto se ponavlja i za uzorak obima n2
M2 <- combn(pop, n2)
psu_2 <- list()
for (i in 1:choose(N, n2)) {
  psu_2[[i]] <- M2[, i]
}
P_uzorka_2 <- rep(1/choose(N, n2), choose(N, n2))

# Za ocene sume trebaju nam uzoračke srednje vrednosti, pa računamo njih prvo
Y_1_sr <- c()
for (i in 1:length(psu_1)) {
  Y_1_sr[i] <- mean(Y[psu_1[[i]]])
}

Y_2_sr <- c()
for (i in 1:length(psu_2)) {
  Y_2_sr[i] <- mean(Y[psu_2[[i]]])
}

# Prava vrednost sume
t <- sum(Y)
# Srednjekvadratne greške
MSE_1 <- sum( (N*Y_1_sr - t)^2 * P_uzorka_1 )
MSE_2 <- sum( (N*Y_2_sr - t)^2 * P_uzorka_2 )
MSE_1
```

```
## [1] 6.666667
```

```r
MSE_2
```

```
## [1] 2.222222
```

Оцена на основу узорка обима 3 има мању средњеквадратну грешку, па је боља.

Задатак 1.4. Из скупа $\{1, 2, 3, 4, 5, 6, 7, 8\}$ ваде се узорци без понављања обима 2, тако да сваки који садржи непаран број има вероватноћу нула. Они узорци који садрже 2 имају вероватноћу обрнуто пропорционалну другом елементу узорка, а они који садрже 4, а не садрже 2, имају вероватноћу $1/8$. Испитати непристрасност оцене средње вредности и одредити њену средњеквадратну грешку.

Решење.

```r
# Populacija
pop <- 1:8
# Obim populacije
N <- length(pop)
# Obeležje je redni broj jedinke
Y <- pop
# Obim uzorka
n <- 2
# Pravimo sve moguće uzorke
M <- combn(pop, n)
uzorci <- list()
```

```r
for (i in 1:choose(N, n)) {
  uzorci[[i]] <- M[, i]
}

# Sada moramo da se domognemo verovatnoća izbora
# Kako drugačije vec naredbama grananja
P_uzorka <- c()
# Napravićemo funkciju koja uzme uzorak i vrati mu verovatnoću
# Sintaksa je slična Matlabovoj
verovatnoca <- function(x) {
  # %% daje ostatak pri deljenju; "%% 2" je nula za parne, a 1 za neparne
  # 1 se tumači kao TRUE, 0 kao FALSE
  if (x[1] %% 2 | x[2] %% 2)
    p = 0
  else if (x[1] == 2 | x[2] == 2)
    p = 1/max(x)
  else if (x[1] == 4 | x[2] == 4)
    p = 1/8 # ako smo ušli u ovu granu, nema dvojki u uzorku
  else
    p = 1 - (1/4 + 1/6 + 1/8 + 1/8 + 1/8) # Uzorak {6, 8} je jedini u ovoj grani
}

# Računamo verovatnoće
for(i in 1:length(uzorci)){
  P_uzorka[i] <- verovatnoca(uzorci[[i]])
}
sum(P_uzorka) # provera da li je zbir verovatnoća = 1
```

```
## [1] 1
```

```r
# Sada tražimo ocenu, već viđeno
Y_sr <- c()
for(i in 1:length(uzorci)){
  Y_sr[i] <- mean(Y[uzorci[[i]]])
}
E_Y_sr <- sum(Y_sr * P_uzorka)
E_Y_sr
```

```
## [1] 4.875
```

```r
# Prava vrednost
mean(Y) # Ocena NIJE nepristrasna
```

```
## [1] 4.5
```

```r
# Računamo jos i srednjekvadratnu grešku
MSE_Y_sr <- sum( (Y_sr - mean(Y))^2 * P_uzorka )
MSE_Y_sr
```

```
## [1] 2.25
```

Задатак 1.5. Из популације $\{1, 2, \ldots, 100\}$ извадити 15 простих случајних узорака без понављања обима 20. На сваком од њих наћи оцену укупне суме обележја и испитати који од узорака је најрепрезентативнији, тј. где је реализована вредност тест статистике $\hat{t} = N\overline{Y}$ најближа стварној вредности.

Решење. Неке делове кода који се већ стално понављају нећемо више тако детаљно коментарисати.

```
populacija <- 1:100
Y <- populacija
n <- 20

# Postavljamo fiksan seed radi ponovljivosti rezultata
set.seed(123)
uzorci <- list()
for (i in 1:15) {
  uzorci[[i]] <- sample(populacija, size = n, replace = FALSE)
}

# Sada imamo izvučenih 15 uzoraka
# Upisujemo ocene na osnovu svakog
t_ocene <- c()
for(i in 1:length(uzorci)) {
  t_ocene[i] = length(populacija) * mean(uzorci[[i]])
}
t_ocene
```

```
##  [1] 5235 5150 5155 4550 6060 5065 6170 5505 4930 4785 4925 4090 5885 5255 5100
```

```
# Suma obeležja na populaciji
t <- sum(populacija)

# Sada cemo da vidimo koja ocena je najbliza pravoj vrednosti t
# To ce biti ona koja ima najmanju apsolutnu razliku sa t
abs_razlike <- abs(t - t_ocene)
# which.min vraća poziciju PRVOG minimuma u vektoru
which.min(abs_razlike)
```

```
## [1] 6
```

```
# Dakle šesti uzorak ima najmanje odstupanje, najreprezentativniji je
# Možemo proveriti i da li ih je više
which(abs_razlike == min(abs_razlike))
```

```
## [1] 6
```

```
# Ipak samo šesti
```

Задатак 1.6. Дата је популација $\{1,2,3,4,5,6,7,8\}$ и размотрен је следећи план узорковања:

S	$\mathbf{P}(S)$
$\{1,3,5,6\}$	$1/8$
$\{2,3,7,8\}$	$1/4$
$\{1,4,6,8\}$	$1/8$
$\{2,4,6,8\}$	$3/8$
$\{4,5,7,8\}$	$1/8$

Испитати непристрасност оцене $\hat{t} = N\overline{Y}$.

Решење.

```
populacija <- 1:8
Y <- populacija  # obeležje = redni broj jedinke
N <- length(populacija)
# Uzorci koji imaju verovatnoću različitu od 0
s1 <- c(1,3,5,6)
s2 <- c(2,3,7,8)
s3 <- c(1,4,6,8)
s4 <- c(2,4,6,8)
s5 <- c(4,5,7,8)
uzorci <- list(s1,s2,s3,s4,s5)
p_uzorka <- c(1/8,1/4,1/8,3/8,1/8)

# t_ocena = N * Y_sr
Y_sr <- c()
for(i in 1:length(uzorci)){
  Y_sr[i] <- mean(Y[uzorci[[i]]])
}
t_ocena <- N * Y_sr
E_t_ocena <- sum(t_ocena * p_uzorka)
E_t_ocena
```

```
## [1] 39.5
```

```
t <- sum(Y)
t # Ocena je pristrasna
```

```
## [1] 36
```

Вежбе 2

ПСУ без понављања: оцене дисперзије, интервали поверења и одређивање обима узорка

На претходним вежбама смо упознали (прост) случајан узорак са и без понављања, а данас ћемо мало детаљаније проучити ПСУ без понављања.

2.1 Неопходно предзнање

Сада ћемо изнети теорију неопходну за праћење задатака са ових вежби. Претпоставићемо да вучемо ПСУ без понављања обима n из популације обима N.

Вероватноћа укључења i-те јединке у узорак једнака је

$$\pi_i = \frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{n}{N}.$$

Вероватноћа укључења i-те и j-те, $i \neq j$, јединке у узорак једнака је

$$\pi_{ij} = \frac{\binom{N-2}{n-2}}{\binom{N}{n}} = \frac{n(n-1)}{N(N-1)}.$$

Следеће што ћемо навести јесу оцене на основу ПСУ без понављања неких типичних параметара, као и њихове особине.

За оцену популацијске средине предложена је следећа оцена:

$$\hat{m}_Y \frac{1}{n} \sum_{k \in S} y_k.$$

Њене особине су:

$$\mathbf{E}\hat{m}_Y = m_Y, \;\; \mathbf{D}\hat{m}_Y = \frac{\sigma^2}{n}\left(1 - \frac{n}{N}\right).$$

Непристрасна оцена ове дисперзије је

$$\hat{\mathbf{D}}\hat{m}_Y = \frac{\overline{S}^2}{n}\left(1 - \frac{n}{N}\right).$$

$100(1-\alpha)\%$ интервал поверења за популацијску средину на основу ПСУ без понављања дат је са

$$\left[\hat{m}_Y - z\sqrt{\frac{\overline{S}^2}{n}\left(1 - \frac{n}{N}\right)}, \;\; \hat{m}_Y + z\sqrt{\frac{\overline{S}^2}{n}\left(1 - \frac{n}{N}\right)}\right],$$

при чему је z квантил реда $1-\alpha/2$ расподеле t_{n-1}. За $n \geq 30$ узимамо да је то квантил нормалне расподеле.

Сличне оцене и њихове особине имамо и за популацијски тотал.

$$\hat{\tau}_Y = N\hat{m}_Y, \;\; \mathbf{E}\hat{\tau}_Y = \tau_Y.$$

$$\mathbf{D}\hat{\tau}_Y = N^2\mathbf{D}\hat{m}_Y, \;\; \hat{D}\hat{\tau}_Y = N^2\hat{\mathbf{D}}\hat{m}_Y.$$

Интервал поверења нивоа $1-\alpha$ дат је са

$$\left[\hat{\tau}_Y - z\sqrt{\frac{N^2\overline{S}^2}{n}\left(1 - \frac{n}{N}\right)}, \;\; \hat{\tau}_Y - z\sqrt{\frac{N^2\overline{S}^2}{n}\left(1 - \frac{n}{N}\right)}\right].$$

2.1.1 Одређивање обима ПСУ без понављања

Када оцењујемо вредност параметра θ оценом $\hat{\theta}$, желимо да та оцена на известан начин буде блиска. Један од честих услова који се од оцене тражи јесте да задовољава

$$\mathbf{P}\{|\hat{\theta} - \theta| > \Delta\} \leq \alpha,$$

за унапред одређене Δ и α. Овај услов заправо значи да је вероватноћа да оцена одступа од праве вредности параметра више од дозвољене толеранције - довољно мала.

Да би оцена $\hat{m}_Y$ код ПСУ без понављања задовољавала овај услов за унапред дате Δ и α, за обим узорка n мора да важи:

$$n \geq \frac{1}{\frac{1}{n_0} + \frac{1}{N}},$$

где је $n_0 = \left(\frac{\sigma \cdot z}{\Delta}\right)^2$, где је z квантил реда $1-\alpha/2$ стандардне нормалне расподеле.

Ако је у питању оцена $\hat{\tau}_Y$, за обим узорка мора да важи

$$n \geq \frac{1}{\frac{1}{n_0} + \frac{1}{N}},$$

где је $n_0 = \left(\frac{N\sigma z}{\Delta}\right)^2$, а z је исти квантил као и малопре.

2.2 Задаци

Задатак 2.1. У ресторану се служе четири врсте колача - баклаве, тулумбе, еклери и шампите. Десеторо људи је купило баклаву, двадесеторо тулумбу, двадесет петоро еклер, а петнаесторо шампиту. На случајан начин се одаберу 2 различите врсте колача и бележи се број људи који су купили ту врсту.

(а) Оценити укупан број особа које су купиле колач, а затим испитати непристрасност оцене која се користи.

(б) Испитати да ли је узорачка дисперзија непристрасна оцена дисперзије на читавој популацији.

Решење.

```
# a)
pop <- 1:4 # populacija: baklava, tulumba, ekler, šampita
Y <- c(10, 20, 25, 15) # obeležje: br. ljudi koji su kupili kolač
N <- length(pop)
n <- 2 # obim uzorka

set.seed(123)
uzorak <- sample(pop, n, replace = F) # izvlačimo PSU bez ponavljanja
t_ocena <- N * mean(Y[uzorak]) # ocena totala na izvučenom uzorku
t_ocena
```

```
## [1] 80
```

```
M <- combn(pop, n)
uzorci <- list() # ovdeređamo sve PSU bez ponavljanja obima n
for(i in 1:choose(N, n)){
  uzorci[[i]] <- M[, i]
}
p_uzorka <- rep(1/6, 6) # svi uzorci su jednako verovatni i imaih 4 nad 2 = 6
Y_sr <- c()
for(i in 1:length(uzorci)){
  Y_sr[i] <- mean(Y[uzorci[[i]]])
}
E_t_ocena <- N * sum(Y_sr*p_uzorka)
t <- sum(Y)
E_t_ocena == t # ocena jeste nepristrasna
```

```
## [1] TRUE
```

```r
# b)

sigma_2 <- var(Y) # populacijska disperzija

S_2 <- c() # uzoračka disperzija
for(i in 1:length(uzorci)){
  S_2[i] <- var(Y[uzorci[[i]]])
}
E_S_2 <- sum(S_2*p_uzorka)

sigma_2 == E_S_2 # ocena jeste nepristrasna
```

```
## [1] TRUE
```

Задатак 2.2. Узет је прост случајан узорак без понављања од 10 ученика од 100 ученика трећег разреда и бележене су њихове оцене из математике. Забележени су резултати $(4, 5, 5, 2, 3, 1, 3, 4, 4, 5)$. Оценити просечну оцену из математике, а затим израчунати оцену дисперзије те оцене.

Решење.

```r
N <- 100
n <- 10
Y_na_uzorku <- c(4, 5, 5, 2, 3, 1, 3, 4, 4, 5) # obeležje su ocene
Y_sr <- mean(Y_na_uzorku) # ocena srednje vrednosti na datom uzorku
Y_sr
```

```
## [1] 3.6
```

```r
ocena_D_Y_sr <- (N - n) * var(Y_na_uzorku) / (N * n) # ocena disperzije ocene Y_sr
# na osnovu datog uzorka
ocena_D_Y_sr
```

```
## [1] 0.164
```

Задатак 2.3. Узет је прост случајан узорак од 10 различитих кућа од 100 кућа које се налазе у једном насељу. Број становника у кућама из узорка је $2, 5, 1, 4, 4, 3, 2, 5, 2, 3$.

(а) Оценити укупан број становника у том насељу и оценити дисперзију те оцене.
(б) Оценити просечан број становника по кући и оценити дисперзију те оцене.
(в) Наћи приближни 90%-ни интервал поверења за укупан број становника.

Решење.

```r
# a)

N <- 100
n <- 10
# obeležje je broj stanovnika u kućama
Y_na_uzorku <- c(2, 5, 1, 4, 4, 3, 2, 5, 2, 3) # vrednosti obeležja na datom uzorku
t_ocena <- N * mean(Y_na_uzorku)
t_ocena # ocena totala na izvučenom uzorku
```

```
## [1] 310
```

```r
ocena_D_t_ocena <- N * (N - n) * var(Y_na_uzorku) / n # ocena disperzije ocene totala
ocena_D_t_ocena
```

```
## [1] 1690
```

```r
# b)

Y_sr <- mean(Y_na_uzorku) # ocena srednje vrednosti na izvučenom uzorku
Y_sr
```

```
## [1] 3.1
```

```r
ocena_D_Y_sr <- (N - n) * var(Y_na_uzorku) / (N * n) # ocena disperzije ocene Y_sr
ocena_D_Y_sr
```

```
## [1] 0.169
```

```r
# c)

alpha <- 1 - 0.9 # želimo 90% interval poverenja
z_student <- qt(1 - alpha/2, n - 1) # n < 30, pa koristimo studentovu raspodelu
I_poverenja_za_t_90 <- c(t_ocena - z_student * sqrt((N - n) * N * var(Y_na_uzorku) / n),
                         t_ocena + z_student * sqrt((N - n) * N * var(Y_na_uzorku) / n))
I_poverenja_za_t_90
```

```
## [1] 234.6414 385.3586
```

Задатак 2.4. Ботаничар жели да оцени број стабала брезе у некој области. Област је подељена на 1000 делова. Познато је из претходних испитивања да је дисперзија броја стабала по области приближно 45. Одредити величину простог случајног узорка без понављања, потребну да са вероватноћом 0.95 одступање не буде веће од 500 стабала.

Решење.

```r
N <- 1000
sigma_2 <- 45
delta <- 500
alpha <- 0.05 # 1 - 0.95
z <- qnorm(1 - alpha/2)
n1 <- (delta^2 / (z^2 * sigma_2 * N^2)+1/N)^{-1}
n <- ceiling(n1)
n
```

```
## [1] 409
```

Задатак 2.5. Посматрамо популацију обима 5, чији су елементи означени бројевима 1, 2, 3, 4 и 5, а вредности обележја су редом 3, 1, 0, 1 и 5. Размотримо принцип простог случајног узорковања без понављања за узорак обима 3. Показати да је средња вредност обележја на узорку непристрасна оцена средње вредности обележја популације.

Решење.

```r
pop <- 1:5
Y <- c(3, 1, 0, 1, 5)
```

```r
N <- length(pop)
n <- 3
M <- combn(pop, n)
uzorci <- list() # ovdeređamo sve PSU bez ponavljanja obima n
for (i in 1:choose(N, n)) {
  uzorci[[i]] <- M[, i]
}
p_uzorka <- rep(1 / choose(N, n), choose(N, n)) # svi uzorci su jednako verovatni i imaih N nad n
Y_sr <- c()
for (i in 1:length(uzorci)) {
  Y_sr[i] <- mean(Y[uzorci[[i]]])
}
E_Y_sr <- sum(Y_sr * p_uzorka)
E_Y_sr == mean(Y) # ocena Y_sr je nepristrasna
```

```
## [1] TRUE
```

Задатак 2.6. У датотеци `deca.txt` дати су подаци о броју деце у свакој од 512 улица у неком граду. Наћи 95%-ни интервал поверења за укупан број деце у том граду користећи прост случајан узорак без понављања обима 200.

```r
deca <- read.table("deca.txt")
head(deca)
```

```
##   ulica br_dece
## 1     1      21
## 2     2      76
## 3     3      50
## 4     4      85
## 5     5      37
## 6     6      63
```

```r
N <- length(deca$ulica)
n <- 200
set.seed(123)
uzorak <- sample(deca$ulica, n, replace = F)
t_ocena <- N * mean(deca$br_dece[uzorak]) # broj dece je obeležje

alpha <- 1 - 0.95 # 95-% interval nas zanima
z <- qnorm(1 - alpha/2) # n = 200 > 30, pa koristimo normalnu raspodelu
S_2 <- var(deca$br_dece[uzorak])
I_poverenja_za_t_95 <- c(t_ocena - z * sqrt(S_2 * N * (N - n) / n),
                         t_ocena + z * sqrt(S_2 * N * (N - n) / n))
I_poverenja_za_t_95
```

```
## [1] 25146.22 28265.62
```

Задатак 2.7. За који од следећих планова простог случајног узорковања без понављања ће бити добијена најпрецизнија оцена средње вредности обележја на популацији?
(а) Узорак обима 400 добијен из популације обима 4000.
(б) Узорак обима 30 добијен из популације обима 300.
(в) Узорак обима 3000 добијен из популације обима 300000000 (три стотине милиона).
Сматрамо да је дисперзија на популацији у сва три случаја једнака 100.

Решење.

```
# tražimo gde je D(Y_sr) najmanja

sigma_2 <- 100

D_Y_sr <- function(N, n) {
  (N - n) * sigma_2 / (N * n)
}
D_Y_sr_1 <- D_Y_sr(4000, 400)
D_Y_sr_2 <- D_Y_sr(300, 30)
D_Y_sr_3 <- D_Y_sr(300000000, 3000)
which.min(c(D_Y_sr_1, D_Y_sr_2, D_Y_sr_3))
```

```
## [1] 3
```

Вежбе 3

Узорковање са неједнаким вероватноћама избора јединки

3.1 Неопходно предзнање

Једна од основних карактеристика простог случајног узорка била је та да су вероватноће избора сваке од јединки у узорак биле међусобно једнаке. Сада ћемо да обрадимо ситуацију у којој то није случај. И у тој ситуацији може се говорити о узорку са и без понављања, а до њих се долази на исти начин: код узорка без понављања јединку не враћамо у популацију након извлачења, а код узорковања са понављањем је враћамо, па је можемо поново извући.

У зависности од тога да ли узорковање вршимо са или без понављања имаћемо различите типове оцена за неке стандардне параметре који се користе.

3.1.1 *Hansen-Hurwitz*-ове оцене

Ове оцене користе се код узорковања *са понављањем*. Може се показати да су извлачење једног узорка једном, или јединку по јединку без враћања - еквивалентне процедуре. Нама је згодније да замислимо да узорак вучемо на овај други начин, јединку по јединку, без враћања. Вероватноћу да i-та јединка из популације буде изабрана у узорак у једном извлачењу означићемо са ψ_i.

Hansen-Hurwitz-ове оцене за популацијски тотал и популацијску средњу вредност, као и њихове особине су следеће:

$$\hat{\tau}_\psi = \frac{1}{n}\sum_{k\in S}\frac{y_k}{\psi_k}, \quad \mathbf{E}\hat{\tau}_\psi = \tau_Y.$$

$$\mathbf{D}\hat{\tau}_\psi = \frac{1}{n}\sum_{k=1}^{N}\psi_k\left(\frac{y_k}{\psi_k}-\tau_Y\right)^2, \quad \hat{\mathbf{D}}\hat{\tau}_\psi = \frac{1}{n(n-1)}\sum_{k\in S}\left(\frac{y_k}{\psi_k}-\hat{\tau}_\psi\right)^2.$$

$$\hat{m}_\psi = \frac{\hat{\tau}_\psi}{N}, \quad \mathbf{E}\hat{m}_\psi = m_Y.$$

$$\mathbf{D}\hat{m}_\psi = \frac{\mathbf{D}\hat{\tau}_\psi}{N^2}, \quad \hat{\mathbf{D}}\hat{m}_\psi = \frac{\hat{D}\hat{\tau}_\psi}{N^2}.$$

3.1.2 *Horvitz-Thompson*-ове оцене

Претходне оцене биле су стриктно за узорковање са понављањем. Како немамо довољно квалитетне, а довољно једноставне оцене за узорковање без понављања, навешћемо тип оцена које су универзалне - користе се и за узорковање са понављањем и без понављања.

Овде ћемо посматрати **вероватноће укључења** π_i и π_{ij}. Подсетимо се, π_i представља вероватноћу да се i-та по реду јединка из популације нађе у извученом узорку, а π_{ij} представља вероватноћу да се у извученом узорку нађу i-та и j-та јединка из популације, за $i \neq j$. Ми ћемо од сад па на даље сматрати да су све ове вероватноће укључења позитивне.

Згодно је навести и наредне везе, јер могу бити корисне за рачун.

$$\pi_k = 1 - (1 - \psi_k)^n.$$
$$\pi_{kl} = \pi_k + \pi_l - 1 + (1 - \psi_k - \psi_l)^n.$$

Нека је сада S' скуп индекса оних јединки које су ушле у узорак S, при чему се индекси јединки које су ушле више пута рачунају само једном. Дакле, специјално ако користимо узорковање без понављања, $S' = S$.

Horvitz-Thompson-ове оцене популацијског тотала, популацијске средње вредности и њихове особине су:

$$\hat{\tau}_\pi = \sum_{k \in S'} \frac{y_k}{\pi_k}, \quad \mathbf{E}\hat{\tau}_\pi = \tau_Y.$$

$$\mathbf{D}\hat{\tau}_\pi = \sum_{k=1}^{N} \frac{1 - \pi_k}{\pi_k} y_k^2 + \sum_{k=1}^{N} \sum_{l \neq k} \frac{\pi_{kl} - \pi_k \pi_l}{\pi_k \pi_l} y_k y_l.$$

$$\hat{\mathbf{D}}\hat{\tau}_\pi = \sum_{k \in S'} \left(\frac{1}{\pi_k^2} - \frac{1}{\pi_k} \right) y_k^2 + \sum_{k \in S'} \sum_{l \neq k} \left(\frac{1}{\pi_k \pi_l} - \frac{1}{\pi_{kl}} \right) y_k y_l.$$

$$\hat{m}_\pi = \frac{\hat{\tau}_\pi}{N}, \quad \mathbf{E}\hat{m}_\pi = m_Y.$$

$$\mathbf{D}\hat{m}_\pi = \frac{\mathbf{D}\hat{\tau}_\pi}{N^2}, \quad \hat{\mathbf{D}}\hat{m}_\pi = \frac{\hat{D}\hat{\tau}_\pi}{N^2}.$$

Може се десити да оцена дисперзије буде негативна. Тада, наравно, она нема смисла.

Горње оцене важе у општем случају: и са и без понављања. Када је узорак баш без понављања, може се добити и мало боље.

Sen-Yates-Grundy-јева оцена дисперзије *Horvitz-Thompson*-ове оцене популацијског тотала за узорковање без понављања дата је са

$$\hat{\mathbf{D}}_{SYG}(\hat{\tau}_\pi) = \frac{1}{2} \sum_{k \in S} \sum_{l \neq k} \frac{\pi_k \pi_l}{\pi_{kl}} \left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l} \right)^2.$$

Испоставља се да *Horvitz-Thompson*-ова оцена, иако непристрасна, може имати велику дисперзију када се примењује на популације код којих π_k и y_k **нису** приближно пропорционални. Тада се прво уведе *Hajek*-ова оцена популацијске средње вредности као:

$$\hat{m}_Y = \frac{\sum_{k \in S'} \frac{y_k}{\pi_k}}{\sum_{k \in S'} \frac{1}{\pi_k}},$$

а затим из ње оцена тотала множењем са N. Видимо да је бројилац класична *Horvitz-Thompson*-ова оцена тотала, док именилац заправо представља непристрасну оцену обима популације N.

3.2 Задаци

Задатак 3.1. Изабран је узорак са вероватноћама пропорционалним величини, са понављањем, обима 3 из популације величине 10. Дате су вредности обележја изабраних елемената и вероватноће избора:

i	y_i	ψ_i
1	3	0.06
2	10	0.2
3	7	0.1

(а) Наћи оцену укупне суме обележја популације користећи *Hansen-Hurwitz*-ову оцену.
(б) Оценити дисперзију те оцене.
(в) Наћи оцену укупне суме обележја популације користећи *Horvitz-Thompson*-ову оцену.
(г) Оценити дисперзију те оцене.

Решење.

```
n <- 3
N <- 10
Y_na_uzorku <- c(3,10,7)
psi <- c(0.06, 0.2, 0.1)

# a)
# Hansen-Hurwits-ova ocena
t_hh <- mean(Y_na_uzorku / psi)
t_hh
```

```
## [1] 56.66667
```

```
# b)
# Ocena disperzije ocene t_hh
ocena_D_t_hh <- sum((Y_na_uzorku / psi - t_hh) ^ 2) / (n * (n - 1))
ocena_D_t_hh
```

```
## [1] 44.44444
```

```
# v)
pi_i <- 1 - (1 - psi) ^ n # verovatnoće uključenja i-tih jedinki
t_ht <- sum(Y_na_uzorku / pi_i) # Horvits-Thompson-ova ocena
t_ht
```

```
## [1] 64.02995
```

```r
# g)
# Ocena disperzije ocene t_ht
ocena_D_t_ht <- sum((1 - pi_i) * Y_na_uzorku ^ 2 / pi_i ^ 2)
for (i in 1:n) {
  for (j in 1:n) {
    if (j != i) {
      pi_ij <-
        pi_i[i] + pi_i[j] - 1 + (1 - psi[i] - psi[j]) ^ n
          # verovatnoća uključenja i-te i j-te jedinke
      ocena_D_t_ht <- ocena_D_t_ht +
        (pi_ij - pi_i[i] * pi_i[j]) * (Y_na_uzorku[i] * Y_na_uzorku[j]) /
        ((pi_i[i] * pi_i[j]) * pi_ij)
    }
  }
}
ocena_D_t_ht
```

```
## [1] 62.4619
```

Задатак 3.2. За испитивање загађености 320 језера укупне површине 80km^2 изабран је узорак са понављањем обима 4 са вероватноћама пропорционалним површини језера. Прво језеро из узорка бирано је два пута, а остала два по једном. Концентрација загађености у та три језера у узорку је редом 2, 5 и 10 милиграма по литру, а површине тих језера у km^2 су редом 1.2, 0.2 и 0.5. Наћи *Hansen-Hurwitz*-ову оцену средњег загађења по језеру у посматраној популацији, као и оцену дисперзије добијене оцене.

Решење.

```r
N <- 320 # jezera su jedinke populacije
n <- 4
M <- 80 # Ukupna površina
uzorak <-
  c(1, 1, 2, 3) # uzorak sa ponavljanjem, obavezno zapisati prvu jedinku dva puta!
Y_na_uzorku <-
  c(2, 2, 5, 10) # obeležje je koncentracija zagađenosti jezera
Mi <-
  c(1.2, 1.2, 0.2, 0.5) # površina jezera će biti njegova "veličina"
psi <- Mi / M
Y_sr_hh <- sum(Y_na_uzorku / psi) / (N * n)
Y_sr_hh
```

```
## [1] 3.020833
```

```r
ocena_D_Y_sr_hh <- sum((Y_na_uzorku / (N * psi) - Y_sr_hh) ^ 2) / (n * (n - 1))
ocena_D_Y_sr_hh
```

```
## [1] 2.325666
```

Задатак 3.3. Из популације коју чине три поља на којима се узгаја пшеница бира се узорак обима 2 са вероватноћама пропорционалним величинама, са понављањем. У следећој табели су дати подаци о количини произведене пшенице на сваком пољу и вероватноће избора сваког поља.

i	y_i	ψ_i
1	11	0.3
2	6	0.2
3	25	0.5

Наћи *Hansen-Hurwitz*-ову и *Horvitz-Thompson*-ову оцену за укупну производњу пшенице за сваки узорак.

Решење.

```
pop <- c(1, 2, 3)
N <- 3 # polja čine populaciju
n <- 2
Y <- c(11, 6, 25) # obeležje: količina proizvedene pšenice
psi <- c(0.3, 0.2, 0.5)

uzorci <- list() # ovde smeštamo sve uzorke sa ponavljanjem obima n
i <- 1
for (j in 1:3) {
  for (k in 1:3) {
    uzorci[[i]] <- c(j, k)
    i <- i + 1
  }
}
length(uzorci) # N^n
```

```
## [1] 9
```

```
t_hh <- c() # računamo t_hh na svakom uzorku
for (i in 1:length(uzorci)) {
  t_hh[i] <- sum(Y[uzorci[[i]]] / psi[uzorci[[i]]]) / n
}
t_hh
```

```
## [1] 36.66667 33.33333 43.33333 33.33333 30.00000 40.00000 43.33333 40.00000
## [9] 50.00000
```

```
pi_i <- 1 - (1 - psi) ^ n
t_ht <- c()
for (i in 1:length(uzorci)) {
  # računamo t_ht na svakom uzorku
  t_ht[i] <- sum(Y[unique(uzorci[[i]])] / pi_i[unique(uzorci[[i]])])
}
t_ht
```

```
## [1] 21.56863 38.23529 54.90196 38.23529 16.66667 50.00000 54.90196 50.00000
## [9] 33.33333
```

Задатак 3.4. У датотеци `radnici.txt` дати су подаци о броју радника и производњи у 10 фабрика у индустријској зони. Изабрати узорак обима 3 са понављањем са вероватноћама избора пропорционалним броју радника у фабрици. Користећи добијени узорак одредити *Hansen-Hurwitz*-ову оцену укупне производње.

Решење.

```
# fabrike su jedinke, a proizvodnja u fabrikama njihovo obeležje
radnici <- read.table("radnici.txt")
head(radnici)
```

```
##   br_radnika proizvodnja
## 1         25        47.3
## 2         30        58.3
## 3         18        27.6
## 4         42        84.7
## 5         11        39.7
## 6         45       101.1
```

```
N <- nrow(radnici)
n <- 3
pop <- 1:N
Mi <- radnici$br_radnika
M <- sum(Mi)
psi <- Mi / M

set.seed(123)
uzorak <- sample(pop, n, replace = T, psi)
uzorak
```

```
## [1] 6 1 4
```

```
t_hh <- mean(radnici$proizvodnja[uzorak]/psi[uzorak])
t_hh
```

```
## [1] 574.4978
```

Задатак 3.5. Дата је популација од четири прашуме, њихове површине и бројеви тигрова који живе у њима:

прашума	површина (у km^2)	број тигрова
1	100	11
2	200	20
3	300	23
4	500	54

Оценити укупан број тигрова, одредити дисперзију те оцене и наћи оцену те дисперзије ако је узорковање вршено са понављањем са вероватноћама пропорционалним површини прашуме. Користити *Hansen-Hurwitz*-ову оцену.

Решење.

```
N <- 4 # prašume čine populaciju
n <- 2
Y <- c(11, 20, 23, 54) # obeležje: broj tigrova
```

```r
Mi <- c(100, 200, 300, 500) # "veličina" jedinke je površina prašume
M <- sum(Mi)
psi <- Mi / M
uzorak <- c(1, 2) # odabrali smo jedan uzorak, reda radi

t_hh <- sum(Y[uzorak] / psi[uzorak]) / n
t_hh
```

```
## [1] 115.5
```

```r
t <- sum(Y)
D_t_hh <- sum(psi * (Y / psi - t) ^ 2) / n
D_t_hh
```

```
## [1] 110.9333
```

```r
ocena_D_t_hh <- sum((Y[uzorak] / psi[uzorak] - t_hh) ^ 2) / (n * (n - 1))
ocena_D_t_hh
```

```
## [1] 30.25
```

Задатак 3.6. Популацију чини база `trees` која садржи податке о 31 дрвету. Изабрати узорак са понављањем обима 10 са вероватноћама пропорционалним обиму стабла (колона `Girth`) и оценити средњу вредност обележја `Volume` користећи *Hansen-Hurwitz*-ову оцену.

.

```r
data("trees") # baza postoji u okviru nekog od paketa u R-u
head(trees)
```

```
##   Girth Height Volume
## 1   8.3     70   10.3
## 2   8.6     65   10.3
## 3   8.8     63   10.2
## 4  10.5     72   16.4
## 5  10.7     81   18.8
## 6  10.8     83   19.7
```

```r
N <- length(trees$Girth)
pop <- 1:N
Y <- trees$Volume
n <- 10
psi <- trees$Girth / sum(trees$Girth)
set.seed(123)
uzorak <- sample(pop, n, replace = T, psi)
Y_sr_hh <- sum(Y[uzorak] / psi[uzorak]) / (N * n)
Y_sr_hh
```

```
## [1] 28.9406
```

Вежбе 4

Количничко оцењивање

4.1 Неопходно предзнање

Количничко оцењивање је техника којом можемо побољшати прецизност оцена посматрајући неко поможно обележје X о коме имамо више информација. Да бисмо користили количничко оцењивање потребно је да веза између обележја Y и помоћног обележја X буде линеарна и да пролази кроз координатни почетак (да нема слободан члан).

Претпостављамо да се вредност обележја X може одредити на произвољној јединки на популацији, као и да знамо тотал τ_X и популацијску средину m_X.

Код количничког оцењивања биће нам значајан и **количник обележја популације**:

$$B = \frac{\tau_Y}{\tau_X} = \frac{m_Y}{m_X}.$$

4.1.1 Количничко оцењивање на основу ПСУ без понављања

Убудуће ћемо вредности обележја X означавати са $x_1, x_2, ...$, а вредности обележја Y са $y_1, y_2, ...$

Посматрајмо сада ПСУ без понављања. **Количничка оцена за параметар** B је тзв. **узорачки количник**:

$$b = \frac{\hat{\tau}_Y}{\hat{\tau}_X} = \frac{\hat{m}_Y}{\hat{m}_X}.$$

Сада ћемо навести неке особине ове оцене, које ће бити изведене на предавањима.

Узорачки количник није непристрасна оцена, али асимптотски јесте. За њену дисперзију и оцену дисперзије важи:

$$\mathbf{D}b \approx \frac{1}{nm_X^2}\left(1 - \frac{n}{N}\right)\frac{1}{N-1}\sum_{i=1}^{N}(y_i - Bx_i)^2.$$

$$\hat{\mathbf{D}}b \approx \frac{1}{nm_X^2}\left(1 - \frac{n}{N}\right)\frac{1}{n-1}\sum_{i\in S}(y_i - bx_i)^2.$$

За количничку оцену популацијске средине и њене особине важи:

$$\hat{m}_Y^r = bm_X.$$

$$\mathbf{D}\hat{m}_Y^r = m_X^2\mathbf{D}b, \;\; \hat{\mathbf{D}}\hat{m}_Y^r = m_X^2\hat{\mathbf{D}}b.$$

За количничку оцену популацијског тотала и њене особине важи:

$$\hat{\tau}_Y^r = b\tau_X.$$

$$\mathbf{D}\hat{\tau}_Y^r = \tau_X^2\mathbf{D}b, \;\; \hat{\mathbf{D}}\hat{\tau}_Y^r = \tau_X^2\hat{\mathbf{D}}b.$$

Ни оцена тотала ни оцена средине нису непристрасне, али јесу асимптотски.

4.1.2 Количничко оцењивање на основу узорковања са неједнаким вероватноћама избора

Оцене које следе користе се и код узорковања са понављањем и код узорковања без понављања. Количничка оцена количника обележја популације је

$$b = \frac{\hat{\tau}_\pi^{(Y)}}{\hat{\tau}_\pi^{(X)}} = \frac{\sum_{i\in S'} \frac{y_i}{\pi_i}}{\sum_{i\in S'} \frac{x_i}{\pi_i}},$$

где су $\hat{\tau}_\pi^{(Y)}$ и $\hat{\tau}_\pi^{(X)}$ *Horvitz-Thompson*-ове оцене тотала ова два обележја.

Оцена b није непристрасна, али јесте асимптотски. Њене особине су

$$\mathbf{D}b \approx \frac{1}{\tau_X^2}\left[\sum_{i=1}^N \frac{1-\pi_i}{\pi_i}(y_i - Bx_i)^2 + \sum_{i=1}^N \sum_{j\neq i} \frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j}(y_i - Bx_i)(y_j - Bx_j)\right].$$

$$\hat{\mathbf{D}}b \approx \frac{1}{\tau_X^2}\left[\sum_{i\in S'}\left(\frac{1}{\pi_i^2} - \frac{1}{\pi_i}\right)(y_i - bx_i)^2 + \sum_{i\in S'}\sum_{j\neq i}\left(\frac{1}{\pi_i\pi_j} - \frac{1}{\pi_{ij}}\right)(y_i - bx_i)(y_j - bx_j)\right].$$

За оцену тотала и популацијске средине важе исте ствари као код ПСУ без понављања:

$$\hat{\tau}_Y^r = b\tau_X.$$

$$\mathbf{D}\hat{\tau}_Y^r = \tau_X^2\mathbf{D}b, \;\; \hat{\mathbf{D}}\hat{\tau}_Y^r = \tau_X^2\hat{\mathbf{D}}b.$$

$$\hat{m}_Y^r = bm_X.$$

$$\mathbf{D}\hat{m}_Y^r = m_X^2\mathbf{D}b, \;\; \hat{\mathbf{D}}\hat{m}_Y^r = m_X^2\hat{\mathbf{D}}b.$$

4.2 Задаци

Задатак 4.1. Узет је прост случајан узорак без понављања обима 30 из велике популације и добијени су подаци $\hat{m}_X = 5$ и $\hat{m}_Y = 18$, где је Y обележје од интереса, а X помоћно обележје. Такође, познато је да је $\tau_X = 1891$. Наћи оцену количника B и количничку оцену суме обележја Y.

Решење.

```
n <- 30
X_sr <- 5
Y_sr <- 18
t_X <- 1891
b <- Y_sr / X_sr
b
```

```
## [1] 3.6
```

```
t_Y_ocena <- b * t_X
t_Y_ocena
```

```
## [1] 6807.6
```

Задатак 4.2. Дата је популација обима 4 и вредности обележја Y које се испитује и помоћног обележја X:

$$X : 3, 5, 7, 9;$$

$$Y : 1, 2, 2, 3.$$

На основу простих случајних узорака без понављања обима 2 и 3 испитати да ли је количничка оцена количника популације непристрасна.

Решење.

```
N <- 4
Y <- c(3, 5, 7, 9) # glavno obeležje
X <- c(1, 2, 2, 3) # pomoćno obeležje
B <- sum(Y) / sum(X) # količnik populacije
B
```

```
## [1] 3
```

```
# uzorci obima 2:
# ima ih 4 nad 2 = 6
X_na_uzorku_obima_2 <- list()
n1 <- 2
for (i in 1:choose(4, 2)) {
  X_na_uzorku_obima_2[[i]] <- combn(X, n1)[, i]
}
X_na_uzorku_obima_2 # Sme da bude ponavljanja! Ovo nisu indeksi.
```

```
## [[1]]
## [1] 1 2
##
## [[2]]
```

```
## [1] 1 2
##
## [[3]]
## [1] 1 3
##
## [[4]]
## [1] 2 2
##
## [[5]]
## [1] 2 3
##
## [[6]]
## [1] 2 3
```

```
Y_na_uzorku_obima_2 <- list()
for (i in 1:choose(4, 2)) {
  Y_na_uzorku_obima_2[[i]] <- combn(Y, n1)[, i]
}

p_uzorka <- rep(1 / choose(4, 2), choose(4, 2))

Y_sr <- c() # ocene srednje vrednosti obeležja Y na svakom od uzoraka
for (i in 1:length(Y_na_uzorku_obima_2)) {
  Y_sr[i] <- mean(Y_na_uzorku_obima_2[[i]])
}
X_sr <- c()
for (i in 1:length(X_na_uzorku_obima_2)) {
  X_sr[i] <- mean(X_na_uzorku_obima_2[[i]])
}
b2 <- Y_sr / X_sr
E_b2 <- sum(b2 * p_uzorka)
E_b2 # E_b2 = B, pa je u ovom slučaju ocena b nepristrasna. Ovo u opštem slučaju ne važi!
```

```
## [1] 3
```

```
X_na_uzorku_obima_3 <- list()
n2 <- 3
for (i in 1:choose(4, 3)) {
  X_na_uzorku_obima_3[[i]] <- combn(X, n2)[, i]
}

Y_na_uzorku_obima_3 <- list()
for (i in 1:choose(4, 3)) {
  Y_na_uzorku_obima_3[[i]] <- combn(Y, n2)[, i]
}

p_uzorka <- rep(1/choose(4,3), choose(4,3))

Y_sr <- c() # ocene srednje vrednosti obeležja Y na svakom od uzoraka
for (i in 1:length(Y_na_uzorku_obima_3)) {
  Y_sr[i] <- mean(Y_na_uzorku_obima_3[[i]])
}
X_sr <- c()
```

```r
for (i in 1:length(X_na_uzorku_obima_3)) {
  X_sr[i] <- mean(X_na_uzorku_obima_3[[i]])
}
b3 <- Y_sr / X_sr
e_b3 <- sum(b3 * p_uzorka)
e_b3 # i ova ocena je nepristrasna
```

```
## [1] 3
```

Задатак 4.3. Дате су следеће вредности:

школа	број деце која имају 5 из математике	број посматраних разреда
1	200	4
2	300	5
3	100	3
4	250	4

Истраживач жели да посматрањем прве и треће школе донесе закључак о укупном броју деце која имају 5 из математике. Испитати да ли је количничка оцена која се користи непристрасна (ако се користи ПСУ без понављања обима 2), одредити њену вредност на датом узорку, а затим упоредити са оценом добијеном на основу ПСУ без понављања истог обима. Која је боља?

Решење.

```r
pop <- c(1, 2, 3, 4) # škole su jedinke populacije
N <- 4
Y <- c(200, 300, 100, 250) # broj dece sa peticom iz matematike je obeležje od interesa
X <- c(4, 5, 3, 4) # broj posmatranih razreda je pomoćno obeležje
uzorak <- pop[c(1, 3)]
t_Y_ocena <- mean(Y[uzorak]) / mean(X[uzorak]) * sum(X) # količnička ocena totala obeležja Y
t_Y_ocena
```

```
## [1] 685.7143
```

```r
N * mean(Y[uzorak]) # ocena totala obeležja Y na osnovu psu bez ponavljanja
```

```
## [1] 600
```

```r
# ispitujemo nepristrasnost količnicke ocene totala na osnovu PSU bez ponavljanja
n <- 2
X_na_uzorku_obima_2 <- list() # ima ih 4 nad 2 = 6
for (i in 1:choose(4, 2)) {
  X_na_uzorku_obima_2[[i]] <- combn(X, n)[, i]
}

Y_na_uzorku_obima_2 <- list()
for (i in 1:choose(4, 2)) {
  Y_na_uzorku_obima_2[[i]] <- combn(Y, n)[, i]
}

p_uzorka <- rep(1 / choose(4, 2), choose(4, 2))

Y_sr <- c() # računamo uzoračku sredinu obeležja Y na svakom uzorku
for (i in 1:length(Y_na_uzorku_obima_2)) {
  Y_sr[i] <- mean(Y_na_uzorku_obima_2[[i]])
```

```
}
X_sr <- c()
for (i in 1:length(X_na_uzorku_obima_2)) {
  X_sr[i] <- mean(X_na_uzorku_obima_2[[i]])
}
b <- Y_sr / X_sr
t_X <- sum(X)
t_Y_ocene <- b * t_X
t_Y_ocene
```

```
## [1] 888.8889 685.7143 900.0000 800.0000 977.7778 800.0000
```

```
E_t_Y_ocena <- sum(t_Y_ocene * p_uzorka)
E_t_Y_ocena
```

```
## [1] 842.0635
```

```
sum(Y) # dakle, ocena nije nepristrasna
```

```
## [1] 850
```

```
# Ocene poredimo na osnovu MSE. Bolja ocena je ona koja ima manju MSE
MSE_kol <- sum((t_Y_ocene - sum(Y)) ^ 2 * p_uzorka)
MSE_kol
```

```
## [1] 8721.55
```

```
sigma_2 <- var(Y) # populacijska disperzija obelezja Y
MSE_psu <- N ^ 2 * (1 - n / N) * sigma_2 / n # MSE nepristrasne ocene je njena disperzija
MSE_psu
```

```
## [1] 29166.67
```

```
MSE_psu > MSE_kol # količnička ocena ima manju MSE
```

```
## [1] TRUE
```

Задатак 4.4. У датотеци `kol.txt` дати су подаци за прост случајан узорак без понављања обима 100 из популације обима 530 у коме је Y обележје које се испитује, а X помоћно обележје. Познато је да је популацијска средња вредност обележја X једнака 3.33.

(а) Наћи количничку оцену средње вредности обележја Y.

(б) Оценити дисперзију добијене количничке оцене.

Решење.

```
kol <- read.table("kol.txt")
head(kol)
```

```
##      y    x
## 1 0.88 5.12
## 2 0.88 6.72
## 3 0.55 0.96
## 4 0.44 6.48
## 5 0.22 7.36
## 6 0.66 4.64
```

```r
N <- 530
n <- 100
Y_na_uzorku <- kol$y
X_na_uzorku <- kol$x

# a)
b <- mean(Y_na_uzorku) / mean(X_na_uzorku)
Y_sr <- b * 3.33 # 3.33 je populacijska sredina obeležja X
Y_sr
```

```
## [1] 0.5194961
```

```r
# b)
# pratimo formulu
ocena_D_Y_sr <- (N - n) * sum((Y_na_uzorku - b * X_na_uzorku) ^ 2) / (N * n * (n - 1))
ocena_D_Y_sr
```

```
## [1] 0.002255486
```

Задатак 4.5. У граду који има 28753 домаћинстава, циљ је да се испита просечан рачун за струју. Посматрана су четири домаћинства и забележен је број чланова, као и рачуни за струју за свако од њих.
Прво домаћинство: 1 члан, рачун од 1282 динара;
Друго домаћинство: 3 члана, рачун од 4375 динара;
Треће домаћинство: 2 члана, рачун од 2333 динара;
Четврто домаћинство: 4 члана, рачун од 5789 динара.
Укупан број становника града је 70351. Наћи количничку оцену просечног рачуна за струју, као и оцену дисперзије те оцене. Колико је укупно новца потрошено на струју?

Решење.

```r
N <- 28753
n <- 4
X_na_uzorku <- c(1, 3, 2, 4) # broj članova domaćinstva je pomoćno obeležje
Y_na_uzorku <- c(1282, 4375, 2333, 5789) # potrošnja novca je glavno obeležje
t_X <- 70351 # populacijski total obeležja X

# a)
b <- mean(Y_na_uzorku) / mean(X_na_uzorku)
Y_sr <- b * t_X / N
Y_sr
```

```
## [1] 3371.358
```

```r
# b)
ocena_D_Y_sr <- (N - n) * sum((Y_na_uzorku - b * X_na_uzorku) ^ 2) / (N * n * (n - 1))
ocena_D_Y_sr
```

```
## [1] 26924.01
```

```r
# c)
t_Y_ocena <- b * t_X
t_Y_ocena
```

```
## [1] 96936643
```

Задатак 4.6. У датотеци `prodavnica.txt` дати су подаци о дневној нето заради за одређени дан у години, подаци о дневној нето заради за исти тај дан претходне године, као и подаци о броју запослених у продавници, за 1534 продавнице у једном насељу. Извадити узорак са понављањем обима 400, са вероватноћама пропорционалним броју запослених, па количничком оценом оценити укупну нето зараду за све продавнице за дан за који се посматра, ако се као помоћно обележје користе зараде од претходне године за исти дан. Израчунати дисперзију те оцене.

Решење.

```
prodavnica <- read.table("prodavnica.txt")
head(prodavnica)
```

```
##     zarada br_radnika zarada_prethodne
## 1 148078.1          5         118462.5
## 2 146593.6          5          73296.8
## 3 156979.6          6         156979.6
## 4 144227.4          5         129804.7
## 5 140885.8          5         140885.8
## 6  50300.9          2          50300.9
```

```
N <- nrow(prodavnica)
pop <- 1:N # populaciju čine prodavnice
Mi <- prodavnica$br_radnika
psi <- Mi / sum(Mi)
n <- 400

# vadimo uzorak
set.seed(123)
uzorak <- sample(pop, n, replace = T, prob = psi)
pi_i <- 1 - (1 - psi) ^ n # verovatnoće uključenja po jedne jedinke

Y <- prodavnica$zarada # glavno obeležje
X <- prodavnica$zarada_prethodne # pomoćno obeležje

uzorak_uq <- unique(uzorak) # za HT ocene nam treba skup S' (S bez duplikata)
t_ht_x <- sum(X[uzorak_uq] / pi_i[uzorak_uq])
t_ht_y <- sum(Y[uzorak_uq] / pi_i[uzorak_uq])
b <- t_ht_y / t_ht_x
t_Y_ocena <- b * sum(X)
t_Y_ocena
```

```
## [1] 248494061
```

```
# Računamo disperziju ocene po formuli
B <- mean(Y) / mean(X)
D_t_ocena <- sum(((1 - pi_i) / pi_i) * (Y - B * X) ^ 2)
for (i in 1:N) {
  for (j in 1:N) {
    if (i != j) {
      pi_ij <- pi_i[i] + pi_i[j] - 1 + (1 - psi[i] - psi[j]) ^ n
```

```
      D_t_ocena <- D_t_ocena +
        ((pi_ij - pi_i[i] * pi_i[j]) / (pi_i[i] * pi_i[j])) * (Y[i] - B * X[i]) * (Y[j] - B * X[j])
    }
  }
}
D_t_ocena
```

```
## [1] 11717017347038
```

Вежбе 5

Регресионо оцењивање

5.1 Неопходно предзнање

Регресионо оцењивање је техника која, попут количничког оцењивања, користи помоћно обележје X за прављење оцена параметара у вези са обележјем Y. И овде претпостављамо да је веза ова два обележја једна права, али овде она **не мора** пролазити кроз координатни почетак. И овде ће за оцењивање бити неопходно знати вредност популацијског тотала τ_X и популацијске средине m_X.

Када у причи имамо (приближно) линеарну везу два обележја, често је од интереса посматрати величину коју називамо **коефицијент корелације**:

$$\rho = \frac{\sum_{i=1}^{N}(y_i - m_Y)(x_i - m_X)}{(N-1)\sigma_Y\sigma_X},$$

где су σ_X и σ_Y стандардна одступања ова два обележја. Што је $|\rho|$ ближе јединици, то је већа линеарна веза ова два обележја.

Када немамо на располагању целу популацију (популације), коефицијент корелације оцењујемо оценом:

$$\hat{\rho} = \frac{\sum_{i\in S}(y_i - \overline{Y})(x_i - \overline{X})}{(n-1)\overline{S}_Y\overline{S}_X},$$

где су у имениоцу (поправљене) узорачке стандардне девијације обележја (корени узорачких дисперзија).

Уколико на основу узорка добијемо да је веза наша два обележја приближно линеарна ($|\hat{\rho}| \approx 1$), можемо користити регресионо оцењивање.

Дакле, ми ћемо претпоставити да за податке важи да је

$$y_i \approx \beta_0 + \beta_1 x_i.$$

5.1.1 Регресионо оцењивање на основу ПСУ без понављања - β_1 непознато

Након што извучемо прост случајан узорак без понављања S, прво питање које се поставља јесте како на основу њега оценити параметре β_0 и β_1. Стандардан начин на који се могу добити те оцене јесте метод најмањих квадрата.

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg\min \sum_{i \in S} (y_i - \beta_0 - \beta_1 x_i)^2.$$

За ове оцене важи:

$$\hat{\beta}_1 = \frac{\sum_{i \in S}(x_i - \overline{X})(y_i - \overline{Y})}{\sum_{i \in S}(x_i - \overline{X})^2}, \quad \hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}.$$

Оцену полулацијске средине обележја Y сада можемо добити као средину оцењених вредности $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$. Добија се да је

$$\hat{m}_Y^{lr} = \overline{Y} + \hat{\beta}_1(m_X - \overline{X}) = \overline{Y} - \hat{\beta}_1(\overline{X} - m_X).$$

Означимо

$$\sigma_d^2 = \frac{1}{N-1} \sum_{k=1}^{N} \left(y_k - (m_Y + \beta_1(x_k - m_X))\right)^2,$$

$$\overline{S}_e^2 = \frac{1}{n-1} \sum_{k \in S} \left(y_k - \left(\overline{Y} + \hat{\beta}_1(x_k - \overline{X})\right)\right)^2.$$

За оцену $\hat{m}_Y^{lr} = \overline{Y} + \hat{\beta}_1(m_X - \overline{X})$ важи да она није непристрасна, а њена пристрасност једнака је $-\mathbf{Cov}(\hat{\beta}_1, \overline{X})$. **Апроксимација** њене дисперзије је

$$\mathbf{D}\hat{m}_Y^{lr} \approx \frac{\sigma_d^2}{n}\left(1 - \frac{n}{N}\right),$$

а она се оцењује са

$$\hat{\mathbf{D}}\hat{m}_Y^{lr} = \frac{\overline{S}_e^2}{n}\left(1 - \frac{n}{N}\right).$$

Оцена тотала формира се као $\hat{\tau}_Y^{lr} = N\hat{m}_Y^{lr}$, а за дисперзију и оцену дисперзије важиће:

$$\mathbf{D}\hat{\tau}_Y^{lr} = N^2\mathbf{D}\hat{m}_Y^{lr}, \quad \hat{\mathbf{D}}\hat{\tau}_Y^{lr} = N^2\hat{\mathbf{D}}\hat{m}_Y^{lr}.$$

5.1.2 Регресионо оцењивање на основу ПСУ без понављања - β_1 познато

У случају да нам је познат коефицијент β_1, оцене и њихове особине незнатно се мењају. За оцену популацијске средине користимо

$$\hat{m}_Y^{lr} = \overline{Y} + \beta_1(m_X - \overline{X}) = \overline{Y} - \beta_1(\overline{X} - m_X).$$

Оцена $\hat{m}_Y^{lr} = \overline{Y} + \beta_1(m_X - \overline{X})$ јесте непристрасна. Њена дисперзија једнака је

$$\mathbf{D}\hat{m}_Y^{lr} = \frac{\sigma_d^2}{n}\left(1 - \frac{n}{N}\right),$$

а она се оцењује са

$$\hat{\mathbf{D}}\hat{m}_Y^{lr} = \frac{\overline{S}_e^2}{n}\left(1 - \frac{n}{N}\right).$$

Оцена тотала формира се као $\hat{\tau}_Y^{lr} = N\hat{m}_Y^{lr}$, а за дисперзију и оцену дисперзије важиће:

$$\mathbf{D}\hat{\tau}_Y^{lr} = N^2\mathbf{D}\hat{m}_Y^{lr}, \quad \hat{\mathbf{D}}\hat{\tau}_Y^{lr} = N^2\hat{\mathbf{D}}\hat{m}_Y^{lr}.$$

Регресионо оцењивање на основу узорковања са неједнаким вероватноћама избора јединки погледати у материјалима са предавања, има само једна формула.

5.2 Задаци

Задатак 5.1. У квалификацијама за европско првенство учествовало је 40 екипа. Одабран је прост случајан узорак без понављања од 6 екипа и дати су подаци о броју постигнутих кошева на 10 утакмица, као и кошаркаша у тиму виших од 2 метра.
Број кошева: $451, 345, 444, 378, 399, 421$.
Виши од 2m: $6, 4, 4, 6, 5, 7$.
Познато је да је укупно на турниру 211 играча виших од 2m. Регресионом методом оценити укупан број постигнутих кошева на десет утакмица, ако је β_1:

(а) Непознат, тј. оцењује се из узорка;
(б) Познат и износи $\beta_1 = 20$. У овом случају наћи и оцену дисперзије оцене укупног броја кошева.

Решење.

```
N <- 40 # ekipe su jedinke
n <- 6
Y_na_uzorku <- c(451, 345, 444, 378, 399, 421) # broj koševa na 10 utakmica je obeležje
                                               # od interesa
X_na_uzorku <- c(6, 4, 4, 6, 5, 7) # broj košarkaša visih od 2m je pomoćno obeležje
t_X <- 211

# a)
beta_1_ocena <-
  sum((X_na_uzorku - mean(X_na_uzorku)) * (Y_na_uzorku - mean(Y_na_uzorku))) /
  sum((X_na_uzorku - mean(X_na_uzorku)) ^ 2)
t_Y_lr <-
  N * (mean(Y_na_uzorku) - beta_1_ocena * (mean(X_na_uzorku) - t_X / N))
t_Y_lr
```

```
## [1] 16231.27
```

```
beta_1 <- 20
t_Y_lr <- N * (mean(Y_na_uzorku) - beta_1 * (mean(X_na_uzorku) - t_X / N))
```

```
t_Y_lr
```

```
## [1] 16206.67
```

```
ocena_D_t_Y_lr <-
  N ^ 2 * (1 - n / N) * (1 / n) * (1 / (n - 1)) * sum((Y_na_uzorku - mean(Y_na_uzorku) -
                                  beta_1 * (X_na_uzorku - mean(X_na_uzorku))) ^ 2)
ocena_D_t_Y_lr
```

```
## [1] 381132.4
```

Задатак 5.2. Истраживач жели да оцени укупан број радно способног становништва у 100 градова. У датотеци `stanovnici.txt` налазе се информације о броју радно способног становништва у 25 градова (претпосавља се да је овај узорак изабран на случајан начин без понављања). Прва колона представља податке прикупљене пре 5 година, а друга податке који се односе на ову годину. Познато је да је пре 5 година било укупно 168740 радно способних становника у ових 100 градова. Наћи регресиону оцену (овогодишњег) укупног броја радно способних становника.

Решење.

```
stanovnici <- read.table("stanovnici.txt")
head(stanovnici)
```

```
##   pre_5_godina ove_godine
## 1           18         25
## 2           30         48
## 3           42         35
## 4          125        180
## 5           73         80
## 6           69         70
```

```
N <- 100 # populaciju čine gradovi
n <- 25 # u uzorak je izabrano 25 gradova
Y_na_uzorku <- stanovnici$ove_godine # glavno obeležje
X_na_uzorku <- stanovnici$pre_5_godina # pomoćno obeležje
t_X <- 168740
beta_1_ocena <-
  sum((X_na_uzorku - mean(X_na_uzorku)) * (Y_na_uzorku - mean(Y_na_uzorku))) /
  sum((X_na_uzorku - mean(X_na_uzorku)) ^ 2)
t_Y_lr <-
  N * (mean(Y_na_uzorku) - beta_1_ocena * (mean(X_na_uzorku) - t_X / N))
t_Y_lr
```

```
## [1] 188453.5
```

Задатак 5.3. Тест из математике на пријемном у једној школи полагало је 486 ученика, који су затим полагали тест из српског. Простим случајним узорковањем без понављања одабрано је 10 студената и забележен је њихов број поена на оба теста. Познато је да просечан број поена на тесту из математике за свих 486 ученика једнак 52. Оценити просечан број поена из српског регресионом оценом, ако су добијени следећи подаци:
Српски: $65, 78, 52, 82, 92, 89, 73, 98, 56, 75$.
Математика: $39, 43, 21, 64, 57, 47, 28, 75, 34, 52$.

Решење.

```
N <- 486 # studenti su jedinke
n <- 10
m_X <- 52
X_na_uzorku <-
  c(39, 43, 21, 64, 57, 47, 28, 75, 34, 52) # poeni iz matematike su pomoćno obeležje
Y_na_uzorku <-
  c(65, 78, 52, 82, 92, 89, 73, 98, 56, 75) # poeni iz srpskog su glavno obeležje
beta_1_ocena <-
  sum((X_na_uzorku - mean(X_na_uzorku)) * (Y_na_uzorku - mean(Y_na_uzorku))) /
  sum((X_na_uzorku - mean(X_na_uzorku)) ^ 2)
Y_sr_lr <-
  mean(Y_na_uzorku) - beta_1_ocena * (mean(X_na_uzorku) - m_X)
Y_sr_lr
```

```
## [1] 80.59337
```

Задатак 5.4. Из пакета `ISLR` учитати базу `Auto` у којој су дате информације о 392 возила. Обележје од интереса је тежина возила (колона `weight`). Проверити да ли има смисла тражити регресиону оцену популацијске средње вредности обележја `weight` преко неког од следећих обележја: `mpg`, `cyllinders`, `displacement`, `horsepower`, `acceleration`.
Извадити ПСУ без понављања обима 100 и наћи регресиону оцену популацијске средње вредности обележја `weight` преко корисних помоћних обележја.

Решење.

```
library(ISLR)
library(corrplot)

data <- Auto
head(Auto)
```

```
##   mpg cylinders displacement horsepower weight acceleration year origin
## 1  18         8          307        130   3504         12.0   70      1
## 2  15         8          350        165   3693         11.5   70      1
## 3  18         8          318        150   3436         11.0   70      1
## 4  16         8          304        150   3433         12.0   70      1
## 5  17         8          302        140   3449         10.5   70      1
## 6  15         8          429        198   4341         10.0   70      1
##                       name
## 1 chevrolet chevelle malibu
## 2         buick skylark 320
## 3        plymouth satellite
## 4             amc rebel sst
## 5               ford torino
## 6          ford galaxie 500
```

```
# sada ćemo ispitati da li postoji linearna zavisnost između weight i ostalih obeležja
corrplot(cor(data[,c("mpg","cylinders", "displacement", "horsepower", "weight","acceleration")]),
         method ="number")
```

Видимо да је `weight` највише корелисана ($|\rho|$ је највеће) са колоном `displacement`, а затим и са колонама `cyllinders`, `horsepower` и `mpg`. Следећи коефицијент корелације је онај са колоном `acceleration` и његова апсолутна вредност је једнака 0.42, што је мало. Ми ћемо се одлучити да нам помоћно обележје буде `displacement`, као оно које је највише корелисано са обележјем од интереса.

> Размислити (не треба за испит) како би се више колона могло користити за побољшање сазнања о обележју од интереса. Да ли би овде имало смисла искористити сва 4 од `cyllinders`, `displacement`, `horsepower`, `mpg`? Какве су њихове међусобне корелације?

```
N <- nrow(data)
pop <- 1:N
Y <- data$weight # glavno obeležje
X <- data$displacement # pomoćno obeležje

set.seed(123)
n <- 100
uzorak <- sample(pop, n, replace = F) # vadimo PSU bez ponavljanja obima n
Y_na_uzorku <- Y[uzorak]
X_na_uzorku <- X[uzorak]

beta_1_ocena <-
  sum((X_na_uzorku - mean(X_na_uzorku)) * (Y_na_uzorku - mean(Y_na_uzorku))) /
  sum((X_na_uzorku - mean(X_na_uzorku)) ^ 2)
```

```
Y_sr_lr <-
  mean(Y_na_uzorku) - beta_1_ocena * (mean(X_na_uzorku) - mean(X))
Y_sr_lr
```

```
## [1] 2964.553
```

```
mean(data$weight) # prava vrednost parametra koji nas interesuje
```

```
## [1] 2977.584
```

```
mean(Y_na_uzorku) # da smo koristili klasičnu ocenu
```

```
## [1] 3126.17
```

Задатак 5.5. На полулацији од 10 јединки дате су вредности главног обележја Y и помоћног обележја X.

$$Y: \ 0.23, 0.74, -0.18, 0.31, 0.14, 0.38, -0.39, 0.90, 0.7, 0.12$$

$$X: \ 0.56, 1.74, -0.33, 0.95, 0.4, 0.93, -0.72, 2.04, 1.59, 0.63.$$

Да ли бисте пре користили регресионо или количничко оцењивање да бисте оценили популацијски тотал τ_Y? Испитати непристрасност количничке и регресионе оцене популацијског тотала обележја Y и упоредити њихове средњеквадратне грешке.

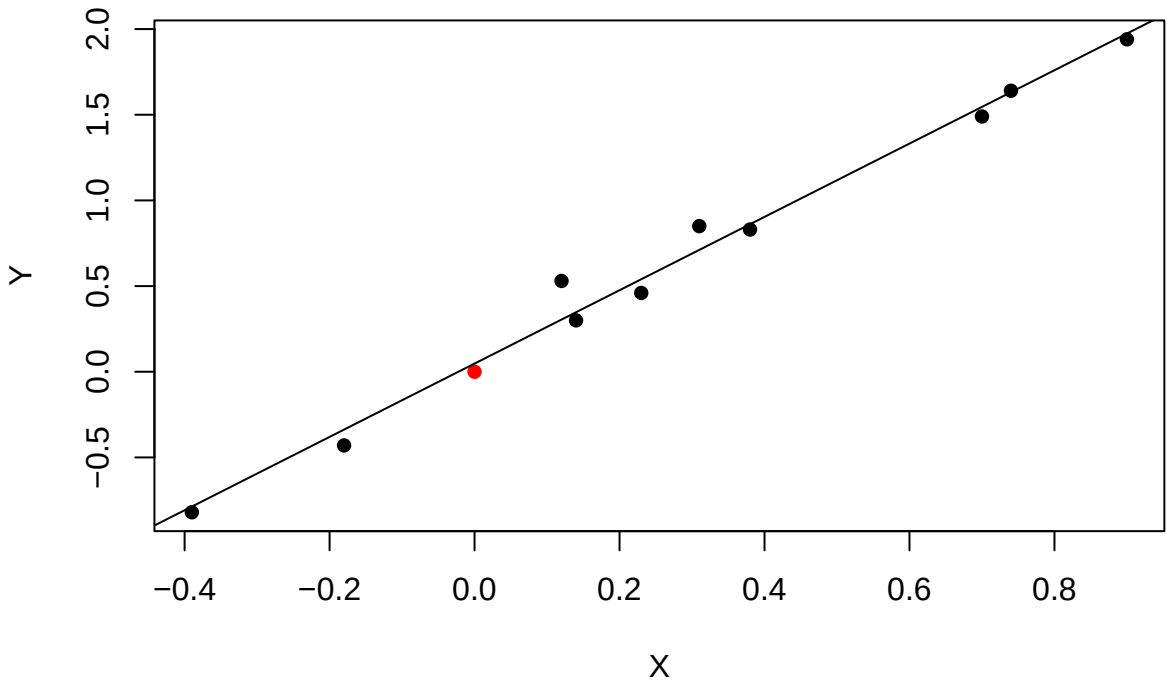
Решење.

```
N <- 10
pop <- 1:N
n <- 2
X <- c(0.23, 0.74, -0.18, 0.31, 0.14, 0.38, -0.39, 0.90, 0.7, 0.12) # pomoćno obeležje
Y <- c(0.46, 1.64, -0.43, 0.85, 0.3, 0.83, -0.82, 1.94, 1.49, 0.53) # glavno obeležje
cor(X, Y)
```

```
## [1] 0.9932019
```

Видимо да је линеарна веза главног и помоћног обележја веома јака (коефицијент корелације је скоро па 1), што значи да можемо користити регресиону оцену. Питање је још да ли можемо користити и количничку оцену. Уколико права зависности пролази кроз координатни почетак онда можемо, иначе не. Ово ћемо утврдити графички.

```
plot(Y ~ X, pch = 16)
points(x = 0, y = 0, pch = 16, col = "red") # koordinatni početak
abline(lm(Y ~ X)) # linearnoregresiona prava
```



Видимо да права коју даје линеарна регресија не одступа превише од координатног почетка, па можемо користити и количничку оцену.

> То што ова права не лежи савршено на црвеној тачки не треба превише да нас брине, јер линеарна веза на подацима никада није савршена, већ увек приближна. Ко жели, може се присетити да је на курсу Статистика рађена линеарна регресија, и да постоје статистички тестови којима се тестира да ли је $\beta_0 = 0$. Овде ће тест заиста показати да је то случај.

Сада прелазимо на други део задатка, испитивање непристрасности оцена и поређење средњеквадратних грешака. То је већ виђен поступак.

```
# Pravimo funkcije koje uzmu uzorak, a vrate odgovarajuću ocenu
t_Y_kol_ocena <- function(uzorak) {
  b <- mean(Y[uzorak]) / mean(X[uzorak])
  t_Y_ocena <- b * sum(X)
  return(t_Y_ocena)
}
t_Y_reg_ocena <- function(uzorak) {
  beta_1_ocena <-
    sum((X[uzorak] - mean(X[uzorak])) * (Y[uzorak] - mean(Y[uzorak]))) / sum((X[uzorak] -
                                                                               mean(X[uzorak])) ^ 2)
  t_Y_ocena <- N * (mean(Y[uzorak]) - beta_1_ocena * (mean(X[uzorak]) - mean(X)))
  return(t_Y_ocena)
}
```

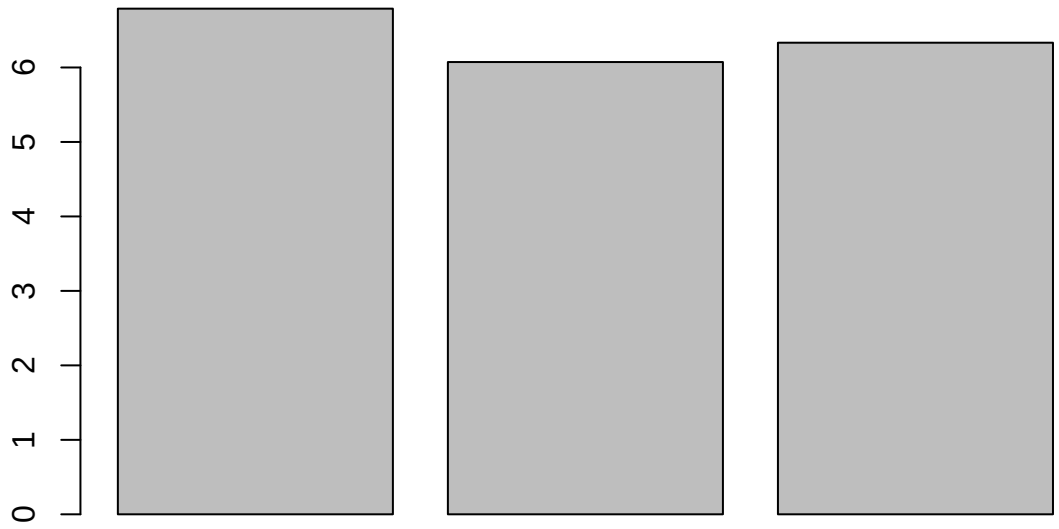
```r
uzorci <- list()
p_uzorka <- rep(1 / choose(N, n), choose(N, n))
t_Y_kol_ocene <- c()
t_Y_reg_ocene <- c()
# sada ćemo za svaki PSU bez ponavljanja izračunati regresionu i količničku
# ocenu populacijskog totala
for(i in 1:choose(N, n)) {
  uzorci[[i]] <- combn(pop, n)[, i]
  t_Y_kol_ocene[i] <- t_Y_kol_ocena(uzorci[[i]])
  t_Y_reg_ocene[i] <- t_Y_reg_ocena(uzorci[[i]])
}
E_t_Y_kol <- sum(t_Y_kol_ocene * p_uzorka)
E_t_Y_kol
```

```
## [1] 6.072818
```

```r
E_t_Y_reg <- sum(t_Y_reg_ocene * p_uzorka)
E_t_Y_reg
```

```
## [1] 6.332562
```

```r
barplot(c(sum(Y), E_t_Y_kol, E_t_Y_reg)) # Obe su pristrasne, manju pristrasnost ima regresiona
```




```r
# Srednjekvadratne greške
MSE_kol <- sum((t_Y_kol_ocene - sum(Y)) ^ 2 * p_uzorka)
```

```
MSE_kol
```

```
## [1] 8.302556
```

```
MSE_reg <- sum((t_Y_reg_ocene - sum(Y)) ^ 2 * p_uzorka)
MSE_reg # manju srednjekvadratnu grešku ima količnička ocena
```

```
## [1] 13.13915
```

У овом задатку регресиона оцена је имала мању пристрасност, али ипак већу средњеквадратну грешку од количничке. Чему је једнака разлика MSE − bias?

Вежбе 6

Стратификован узорак

6.1 Неопходно предзнање

Идеја стратификованог узорка јесте да се популација подели на међусобно дисјунктне скупове, тако да су јединке унутар једног од тих скупова међусобно сличне (у смислу посматраног обележја), док су јединке из различитих скупова међусобно различите. Такви скупови зову се **стратуми**, па отуда и име узорка - стратификован узорак.

Претпоставимо да је популација подељена на L стратума. Из сваког од стратума вадићемо узорак по жељеном плану узорковања, и тако ћемо добити: стратификован случајан узорак без понављања, стратификован случајан узорак са понављањем, итд.

Захтевамо да су стратуми дисјунктни и да покривају целу популацију. Такође, захтевамо и да се у сваком стратуму нађу барем две јединке, како би постојала могућност случајног одабира узорка из стратума. Сада ћемо увести неке стандардне ознаке које се користе при стратификованом узорковању:

- L - укупан број стратума;
- N_h - број јединки у h-том стратуму, $h \in \{1, 2, \dots, L\}$;
- y_{hi} - вредност обележја i-те јединке h-тог стратума, $i \in \{1, 2, \dots, N_h\}$;
- n_h - број јединки које се бирају у узорак из h-тог стратума;
- $\tau_h = \sum_{i=1}^{N_h} y_{hi}$ - тотал h-тог стратума;
- $m_h = \frac{1}{N_h} \sum_{i=1}^{N_h} y_{hi}$ - средина h-тог стратума;
- $\sigma_h^2 = \frac{1}{N_h - 1} \sum_{i=1}^{N_h} (y_{hi} - m_h)^2$ - дисперзија h-тог стратума.

Задржавајући старе ознаке: N за величину популације, n за величину узорка и τ_Y за популацијски тотал, имамо да важи да је

$$N = \sum_{h=1}^{L} N_h, \quad n = \sum_{h=1}^{L} n_h, \quad \tau_Y = \sum_{h=1}^{L} \tau_h.$$

6.1.1 Стратификован случајан узорак

Стратификован случајан узорак добијамо тако што из сваког стратума вадимо прост случајан узорак, при чему су вађења из различитих стратума независна. Јасно је да у овом случају можемо разликовати

стратификован случајан узорак без понављања и са понављања. Означимо са S_h скуп јединки које су ушле у узорак из h-тог стратума, а са $\overline{S}_h^2$ (поправљену) узорачку дисперзију узорка из h-тог стратума.

За стратификован случајан узорак **без понављања** непристрасна оцена популацијског тотала дата је са

$$\hat{\tau}_Y^{str} = \sum_{h=1}^{L} N_h \hat{m}_h = \sum_{h=1}^{L} N_h \frac{1}{n_h} \sum_{k \in S_h} y_{hk}.$$

За дисперзију ове оцене важи да је

$$\mathbf{D}\hat{\tau}_Y^{str} = \sum_{h=1}^{L} N_h^2 \frac{\sigma_h^2}{n} \left(1 - \frac{n_h}{N_h}\right),$$

а непристрасна оцена ове дисперзије је

$$\hat{\mathbf{D}}\hat{\tau}_Y^{str} = \sum_{h=1}^{L} N_h^2 \frac{\overline{S}_h^2}{n} \left(1 - \frac{n_h}{N_h}\right).$$

Непристрасна оцена популацијске средине и њене особине дате су са:

$$\hat{m}_Y^{str} = \frac{\hat{\tau}_Y^{str}}{N}, \quad \mathbf{D}\hat{m}_Y^{str} = \frac{\mathbf{D}\hat{\tau}_Y^{str}}{N^2}, \quad \hat{\mathbf{D}}\hat{m}_Y^{str} = \frac{\hat{\mathbf{D}}\hat{\tau}_Y^{str}}{N^2}.$$

За стратификован случајан узорак **са понављањем** непристрасна оцена популацијског тотала дата је са

$$\hat{\tau}_Y^{str} = \sum_{h=1}^{L} N_h \hat{m}_h.$$

Дисперзија те оцене једнака је

$$\mathbf{D}\hat{\tau}_Y^{str} = \sum_{h=1}^{L} \frac{N_h(N_h - 1)\sigma_h^2}{n_h}.$$

Непристрасна оцена те дисперзије је

$$\hat{\mathbf{D}}\hat{\tau}_Y^{str} = \sum_{h=1}^{L} \frac{N_h^2 \overline{S}_h^2}{n_h}.$$

Оцена популацијске средине, као и њена дисперзија, добијају се на начин аналоган оном у ССУ без понављања.

6.1.2 Одређивање обима ССУ без понављања по стратумима

Нека имамо стратификовано случајно узорковање. Природно се намеће питање како изабрати $n_1, n_2, \ldots, n_L$ тако да оцена буде што боља. У зависности од тога шта од оцене тражимо, истичу се различити типови одабира обима.

Пропорционални избор обима

Код овог метода број јединки које се бирају у узорак из сваког стратума је пропорционалан броју јединки у стратуму:

$$\frac{n_1}{N_1} = \frac{n_2}{N_2} = \cdots = \frac{n_L}{N_L} = \frac{n}{N},$$

одакле видимо да је $n_h = \frac{n}{N} N_h$. Јасно, често ће се десити да ово није цео број, те ћемо тада заокругљивати, пазећи да у збиру и даље добијемо жељени обим.

Није тешко уочити да се у овом случају оцене популацијске средине и популацијског тотала рачунају као и код ПСУ без понављања.

Овај метод пожељно је користити када је дисперзија обележја по стратумима приближно једнака.

Оптимални избор обима

Извлачење узорка, генерално говорећи, изискује неке трошкове. Са c_0 ћемо означити трошкове истраживања који су стални и не зависе од тога колико јединки и одакле узоркујемо. Са c_h означићемо трошак по јединки из h-тог стратума.

Тада се укупни трошкови при целом процесу узорковања могу изразити као

$$C = c_0 + \sum_{h=1}^{L} c_h n_h.$$

Циљ наше оцене је да има што мању дисперзију. Дакле, ми заправо решавамо оптимизациони проблем

$$\min_{(n_1,\ldots,n_h)} \mathbf{D}\hat{\tau}_Y^{str}, \text{ при услову } C = c_0 + \sum_{h=1}^{L} c_h n_h.$$

Другим речима, унапред знамо колико новца треба да потрошимо на истраживање и са тим новцем треба да „извучемо'' најбољу могућу оцену.

Може се показати да се при овој минимизацији добију вредности

$$n_h = \frac{(C - c_0)\frac{N_h \sigma_h}{\sqrt{c_h}}}{\sum_{k=1}^{L} \sqrt{c_k} N_k \sigma_k}, \quad h \in \{1, 2, \ldots, L\}.$$

Укупан обим узорка који овде користимо јесте $n = \sum_{h=1}^{L} n_h$.

Оптимални Нејманов избор обима

Ако претпоставимо да су трошкови по јединки једнаки за сваки стратум, односно да је $c_1 = c_2 = \cdots = c_L = c$, претходни метод се своди на налажење $n_1, \ldots, n_L$ минимизацијом дисперзије оцене тотала када нам је обим узорка унапред дат и износи $n = n_1 + \cdots + n_L$.

У овој ситуацији добије се

$$n_h = n \cdot \frac{N_h \sigma_h}{\sum_{k=1}^{L} N_k \sigma_k}, \quad h \in \{1, 2, \ldots, L\}.$$

6.2 Задаци

Задатак 6.1. Популација од 6 јединки подељена је на 2 стратума, тако да се у првом налазе јединке са вредностима обележја 0, 1 и 2, а у другом јединке са вредностима обележја 4, 6 и 11. У стратификован узорак су методом простог случајног узорковања без понављања изабране јединке са вредностима обележја 0 и 2 из првог стратума, а 6 и 11 из другог.

(а) Одредити оцену средње вредности обележја популације, дисперзију те оцене и оцену те дисперзије.

(б) Упоредити добијену оцену са оценом средње вредности предложене код простог случајног узорковања без понављања.

(в) Оценити укупну суму обележја популације, дисперзију те оцене и оцену те дисперзије на основу добијеног стратификованог узорка.

Решење.

```
N <- 6
Y_str1 <- c(0, 1, 2)
Y_str2 <- c(4, 6, 11)
N1 <- N2 <- 3
Y_na_uzorku_iz_str1 <- c(0, 2)
Y_na_uzorku_iz_str2 <- c(6, 11)
n1 <- 2
n2 <- 2
n <- 4

# a)
mY_str <- (N1 * mean(Y_na_uzorku_iz_str1) + N2 * mean(Y_na_uzorku_iz_str2)) / N
mY_str
```

```
## [1] 4.75
```

```
sigma_2_str1 <- var(Y_str1)
sigma_2_str2 <- var(Y_str2)
D_mY_str <- (N1 ^ 2 * sigma_2_str1 / n1 * (1 - n1 / N1) + N2 ^ 2 * sigma_2_str2 / n2 *
     (1 - n2 / N2)) / N ^ 2
D_mY_str
```

```
## [1] 0.5833333
```

```
S_2_str1 <- var(Y_na_uzorku_iz_str1)
S_2_str2 <- var(Y_na_uzorku_iz_str2)
ocena_D_mY_str <-
  (N1 ^ 2 * S_2_str1 / n1 * (1 - n1 / N1) + N2 ^ 2 * S_2_str2 / n2 * (1 - n2 / N2)) / N ^ 2
ocena_D_mY_str
```

```
## [1] 0.6041667
```

```r
# b)
# Znamo da ocene poredimo u srednjekvaratnom smislu. Kako se i stratifikovana ocena
# i ocena na osnovu PSU nepristrasne, njihove srednjekvadratne greške su jednake
# njihovoj disperziji.

sigma_2 <- var(c(Y_str1,Y_str2))
D_Y_sr <- sigma_2/n*(1-n/N) # disperzija ocene populacijske sredine na osnovu PSU
D_Y_sr
```

```
## [1] 1.366667
```

```r
D_Y_sr > D_mY_str # dakle, stratifikovana ocena je bolja jer ima manju MSE
```

```
## [1] TRUE
```

```r
# v)
t_str <- N * mY_str
t_str
```

```
## [1] 28.5
```

```r
D_t_str <- N^2*D_mY_str
D_t_str
```

```
## [1] 21
```

```r
ocena_D_t_str <- N ^ 2 * ocena_D_mY_str
ocena_D_t_str
```

```
## [1] 21.75
```

Задатак 6.2. Испитује се број претрчаних метара ученика једног разреда на часу физичког васпитања. У разреду има 112 ученика, од чега 59 девојчица и 53 дечака. Посматрано је 7 дечака и 8 девојчица.
Дечаци: $879, 810, 789, 567, 900, 870, 777$.
Девојчице: $450, 234, 679, 456, 239, 555, 560, 467$.

(а) Коришћењем стратификованог узорка у коме су ученици разврстани у стратуме према полу оценити укупан број метара које су претрчала деца тог разреда и оценити дисперзију те оцене. Затим урадити исто коришћењем оцена предложених код простог случајног узорковања без понављања.

(б) Оценити просечан број претрчаних метара и оценити дисперзију те оцене, такође коришћењем стратификованог узорка.

Решење.

```r
N <- 112
# devojčice i dečaci su u dva različita stratuma
N_decaka <- 53
N_devojcica <- 59
uzorak_decaci <- c(879, 810, 789, 567, 900, 870, 777)
uzorak_devojcice <- c(450, 234, 679, 456, 239, 555, 560, 467)
n_decaka <- 7
n_devojcica  <- 8
n <- 15
```

```r
# a)
t_str <-
  (N_decaka * mean(uzorak_decaci) + N_devojcica * mean(uzorak_devojcice))
t_str
```

```
## [1] 69184.43
```

```r
S_2_decaci <- var(uzorak_decaci)
S_2_devojcice <- var(uzorak_devojcice)
ocena_D_t_str <-
  N_decaka ^ 2 * S_2_decaci / n_decaka * (1 - n_decaka / N_decaka) +
  N_devojcica ^ 2 * S_2_devojcice / n_devojcica * (1 - n_devojcica / N_devojcica)
ocena_D_t_str
```

```
## [1] 13387712
```

```r
# ocena totala na osnovu PSU
t_ocena <- N * mean(c(uzorak_decaci, uzorak_devojcice))
t_ocena
```

```
## [1] 68932.27
```

```r
S_2 <- var(c(uzorak_decaci, uzorak_devojcice))
ocena_D_t_ocena <- N ^ 2 * S_2 / n * (1 - n / N)
ocena_D_t_ocena
```

```
## [1] 35409384
```

```r
# b)
mY_str <- t_str / N
mY_str
```

```
## [1] 617.7181
```

```r
ocena_D_mY_str <- ocena_D_t_str / N ^ 2
ocena_D_mY_str
```

```
## [1] 1067.26
```

Задатак 6.3. Популација је подељена на три стратума чије су величине $N_1 = 123$, $N_2 = 102$ и $N_3 = 180$ и одговарајуће дисперзије обележја су $\sigma_1^2 = 116$, $\sigma_2^2 = 143$ и $\sigma_3^2 = 170$.

(а) Ако се бира стратификован узорак обима 10, одредити величину узорка који се вади из сваког стратума користећи пропорционални избор узорка.

(б) Ако се бира стратификован узорак обима 10, одредити величину узорка који се вади из сваког стратума користећи Нејманов оптимални избор узорка.

Решење.

```r
Ni <- c(123, 102, 180)
sigma_2_i <- c(116, 143, 170)
n <- 10
N <- sum(Ni)

# a)
# proporcionalni izbor
```

```r
n_prop <- round(n * Ni / N)
n_prop
```

```
## [1] 3 3 4
```

```r
sum(n_prop) # mora biti jednaka n, u suprotnom neki od n1,n2 i n3 smanjimo ili povećamo
```

```
## [1] 10
```

```r
# b)
# Nejmanov izbor
n_Nejman <- n * Ni * sqrt(sigma_2_i) / sum(Ni * sqrt(sigma_2_i))
n_Nejman <- round(n_Nejman)
n_Nejman
```

```
## [1] 3 2 5
```

```r
sum(n_Nejman) # opet proverimo, dobro je
```

```
## [1] 10
```

Задатак 6.4. За испитивање просечне недељне потрошње бензина град је подељен на 4 дела, који се посматрају као стратуми. Извађен је стратификован случајан узорак (без понављања) и забележена је потрошња бензина (у литрима) за протеклу недељу код сваког возача из узорка. Добијени су следећи подаци:
Стратум 1: $N_1 = 3750$, $\hat{m}_1 = 12.6$;
Стратум 2: $N_2 = 3272$, $\hat{m}_2 = 14.5$;
Стратум 3: $N_3 = 1387$, $\hat{m}_3 = 18.6$;
Стратум 4: $N_4 = 2475$, $\hat{m}_4 = 13.8$.

(а) Оценити просечну недељну потрошњу за цео град.

(б) Ако треба изабрати стратификован случајан узорак обима 1000, одредити величину узорка који се вади из сваког стратума користећи пропорционални избор.

Решење.

```r
Ni <- c(3750, 3272, 1387, 2475)
Y_sr_i <- c(12.6, 14.5, 18.6, 13.8)

# a)
N <- sum(Ni)
mY_str <- sum(Ni * Y_sr_i) / N
mY_str
```

```
## [1] 14.20867
```

```r
# b)
n <- 1000
n_prop <- round(n*Ni/N)
n_prop
```

```
## [1] 345 301 127 227
```

```r
sum(n_prop) # provera
```

```
## [1] 1000
```

Задатак 6.5. Средњошколци су подељени у три групе по успеху у школи са циљем да се испита колико су заинтересовани за позориште. У фајлу `posete.txt` дати су подаци о броју посета позоришту за ученике током годину дана, као и којој групи по успеху припадају. Изабрати стратификовани узорак без понављања обима 200 користећи пропорционални избор, а затим оценити просечан број посета позоришту током годину дана за средњошколце, дисперзију те оцене и оцену те дисперзије.

```
posete <- read.table("posete.txt")
head(posete)
```

```
##   grupa br_poseta
## 1     1        23
## 2     2        28
## 3     1         0
## 4     3        14
## 5     2        29
## 6     3        21
```

```
n <- 200
attach(posete) # ne moramo više kolonama pristupati koristeći simbol $
stratumi <- list()
stratumi[[1]] <- posete[grupa == 1, ]
stratumi[[2]] <- posete[grupa == 2, ]
stratumi[[3]] <- posete[grupa == 3, ]

Ni <-
  c(sum(grupa == 1), sum(grupa == 2), sum(grupa == 3)) # obimi stratuma
Ni
```

```
## [1] 135 143 122
```

```
N <- sum(Ni) # ili length(grupa)
N
```

```
## [1] 400
```

```
n_prop <- round(n * Ni / N) # proporcionalni izbor
n_prop
```

```
## [1] 68 72 61
```

```
sum(n_prop) # veće od n, moramo neki ni da smanjimo
```

```
## [1] 201
```

```
i <- sample(1:3, 1)
n_prop[i] <- n_prop[i] - 1
sum(n_prop)
```

```
## [1] 200
```

```
uzorak_prop <- list()
set.seed(123)
# broj poesta je obeležje
uzorak_prop[[1]] <-
```

```
  sample(stratumi[[1]]$br_poseta, n_prop[1], replace = F)
uzorak_prop[[2]] <-
  sample(stratumi[[2]]$br_poseta, n_prop[2], replace = F)
uzorak_prop[[3]] <-
  sample(stratumi[[3]]$br_poseta, n_prop[3], replace = F)

# disperzije obeležja na stratumima
sigma_2_i <-
  c(var(stratumi[[1]]$br_poseta),
    var(stratumi[[2]]$br_poseta),
    var(stratumi[[3]]$br_poseta))

# uzoracke disperzije po stratumima za odabrani uzorak
S_2_prop <-
  c(var(uzorak_prop[[1]]), var(uzorak_prop[[2]]), var(uzorak_prop[[3]]))

mY_str_prop <-
  (Ni[1] * mean(uzorak_prop[[1]]) + Ni[2] * mean(uzorak_prop[[2]])
   + Ni[3] * mean(uzorak_prop[[3]])) / N
mY_str_prop
```

```
## [1] 16.93141
```

```
# Racunamo disperziju
D_mY_str_prop <-
  (
    Ni[1] ^ 2 * sigma_2_i[1] / n_prop[1] * (1 - n_prop[1] / Ni[1])
    + Ni[2] ^ 2 * sigma_2_i[2] / n_prop[2] * (1 - n_prop[2] / Ni[2])
    + Ni[3] ^ 2 * sigma_2_i[3] / n_prop[3] * (1 - n_prop[3] / Ni[3])
  ) / N ^ 2
D_mY_str_prop
```

```
## [1] 0.2364169
```

```
# Racunamo ocenu disperzije
ocena_D_mY_str_prop <-
  (
    Ni[1] ^ 2 * S_2_prop[1] / n_prop[1] * (1 - n_prop[1] / Ni[1])
    + Ni[2] ^ 2 * S_2_prop[2] / n_prop[2] * (1 - n_prop[2] / Ni[2])
    + Ni[3] ^ 2 * S_2_prop[3] / n_prop[3] * (1 - n_prop[3] / Ni[3])
  ) / N ^ 2
ocena_D_mY_str_prop
```

```
## [1] 0.2268963
```

Вежбе 7

Кластер узорак

7.1 Неопходно предзнање

Насупрот стратификованом узорку, у којем смо се трудили да стратуми унутар себе буду хомогени, а међусобно хетерогени, код кластер узорка популацију ћемо испарчати на **кластере**, тако да сваки од кластера што репрезентативније осликава целу популацију.

Кластере ћемо звати примарним јединкама, а њихове елементе секундарним јединкама. На случајан начин бираћемо примарне јединке, а у узорак ће ући све секундарне јединке које се налазе у оквиру изабране примарне јединке.

Ознаке које ћемо користити код овог типа узорковања знатно се разликују од ознака код других типова узорковања, те се саветује опрез. Са N ћемо означавати укупан број **кластера**, а са M_l, $l \in \{1, 2, \dots, N\}$ број секундарних јединки у l-том кластеру. Обим целе популације (број секундарних јединки) означићемо са M; важи, дакле, да је

$$M = \sum_{l=1}^{N} M_l.$$

Број кластера чије ће секундарне јединке ући у узорак означићемо са n. Остале ознаке које ћемо користити су:

- y_{lk} - вредност обележја Y на k-тој јединки l-тог кластера;
- $\tau_l = \sum_{k=1}^{M_l} y_{lk}$ - тотал l-тог кластера;
- $m_l = \frac{\tau_l}{M_l}$ - средина l-тог кластера;
- $\sigma_l^2 = \frac{1}{M_l - 1} \sum_{k=1}^{M_l} (y_{lk} - m_l)^2$ - дисперзија l-тог кластера.

Није тешко уочити да за популацијски тотал и популацијску средину (секундарних јединки) важи да је

$$\tau_Y = \sum_{l=1}^{N} \tau_l, \quad m_Y = \frac{\tau_Y}{M}.$$

Сада ћемо кластере посматрати као јединке нове популације, а тотале/средине кластера као ново обележје.

Остало је још да изложимо пар планова узорковања по којима ћемо бирати кластере у узорак.

7.1.1 Кластер узорак код којег се примарне јединке бирају као ПСУ без понављања

Овакав узорак састоји се од јединки n кластера одабраних од свих N кластера као ПСУ без понављања.

При ПСУБП одабиру кластера оцене параметара τ_Y и m_Y и њихове особине дате су са:

$$\hat{\tau}_Y^u = \frac{N}{n}\sum_{l\in S}\tau_l, \;\; \mathbf{E}\hat{\tau}_Y^u = \tau_Y.$$

$$\mathbf{D}\hat{\tau}_Y^u = \frac{N^2}{n}\left(1-\frac{n}{N}\right)\frac{1}{N-1}\sum_{l=1}^{N}\left(\tau_l - \frac{\tau_Y}{N}\right)^2, \;\; \hat{\mathbf{D}}\hat{\tau}_Y^u = \frac{N^2}{n}\left(1-\frac{n}{N}\right)\frac{1}{n-1}\sum_{l\in S}\left(\tau_l - \frac{1}{n}\sum_{k\in S}\tau_k\right)^2.$$

$$\hat{m}_Y^u = \frac{\hat{\tau}_Y^u}{M}, \;\; ,\mathbf{E}\hat{m}_Y^u = m_Y.$$

$$\mathbf{D}\hat{m}_Y^u = \frac{\mathbf{D}\hat{\tau}_Y^u}{M^2}, \;\; \hat{\mathbf{D}}\hat{m}_Y^u = \frac{\hat{\mathbf{D}}\hat{\tau}_Y^u}{M^2}$$

7.1.2 Кластер узорак код кога се примарне јединице бирају са вероватноћама пропорционалним величинама кластера

Овакав узорак се састоји од јединки оних n кластера изабраних од свих N кластера методом узорковања са неједнаким вероватноћама избора; при томе је вероватноћа избора једног кластера пропорционална његовој величини, односно:

$$\psi_l = \frac{M_l}{M}, \;\; l \in \{1, 2, \ldots, N\}.$$

За овакав узорак ћемо, као и раније, разликовати оцене добијене на основу узорковања кластера са и без понављања.

Hansen-Hurwitz-ове оцене

Ове оцене користимо када смо кластере извлачили са понављањем. Оцене и њихове особине дате су испод.

$$\hat{\tau}_Y^\psi = \frac{M}{n}\sum_{l\in S}\frac{\tau_l}{M_l}, \;\; \mathbf{E}\hat{\tau}_Y^\psi = \tau_Y.$$

$$\mathbf{D}\hat{\tau}_Y^\psi = \frac{M}{n}\sum_{l=1}^{N}M_l\left(\frac{\tau_l}{M_l} - \frac{\tau_Y}{M}\right)^2, \;\; \hat{\mathbf{D}}\hat{\tau}_Y^\psi = \frac{1}{n(n-1)}\sum_{l\in S}M_l\left(\frac{\tau_l}{M_l}M - \hat{\tau}_Y^\psi\right)^2.$$

$$\hat{m}_Y^\psi = \frac{\hat{\tau}_Y^\psi}{M}, \;\; \mathbf{E}\hat{m}_Y^\psi = m_Y.$$

$$\mathbf{D}\hat{m}_Y^\psi = \frac{\mathbf{D}\hat{\tau}_Y^\psi}{M^2}, \;\; \hat{\mathbf{D}}\hat{m}_Y^\psi = \frac{\hat{\mathbf{D}}\hat{\tau}_Y^\psi}{M^2}.$$

Horvitz-Thompson-ове оцене

Ове оцене користимо и код извлачења кластера са понављањем и без понављања. Оцене и њихове особине у потпуној су аналогији са онима где смо узорковали (секундарне) јединке. Уколико са S' означимо узорак кластера без дупликата, а са π_l и π_{lk} вероватноће укључења l-тог, односно l-тог и k-тог кластера у узорак, добијамо оцене, дисперзије и непристрасне оцене дисперзија дате формулама:

$$\hat{\tau}_Y^\pi = \sum_{l \in S'} \frac{\tau_l}{\pi_l}, \quad \mathbf{E}\hat{\tau}_Y^\pi = \tau_Y.$$

$$\mathbf{D}\hat{\tau}_Y^\pi = \sum_{l=1}^{N} \frac{1-\pi_l}{\pi_l}\tau_l^2 + \sum_{l=1}^{N}\sum_{k \neq l} \frac{\pi_{lk}-\pi_l\pi_l}{\pi_l\pi_k}\tau_l\tau_k.$$

$$\hat{\mathbf{D}}\hat{\tau}_Y^\pi = \sum_{l \in S'} \left(\frac{1}{\pi_l^2} - \frac{1}{\pi_l}\right)\tau_l^2 + \sum_{l \in S'}\sum_{k \neq l} \left(\frac{1}{\pi_l\pi_k} - \frac{1}{\pi_{lk}}\tau_k\tau_k\right).$$

$$\hat{m}_Y^\pi = \frac{\hat{\tau}_Y^\pi}{M}, \quad \mathbf{E}\hat{m}_Y^\pi = m_Y.$$

$$\mathbf{D}\hat{m}_Y^\pi = \frac{\mathbf{D}\hat{\tau}_Y^\pi}{M^2}, \quad \hat{\mathbf{D}}\hat{m}_Y^\pi = \frac{\hat{\mathbf{D}}\hat{\tau}_Y^\pi}{M^2}.$$

7.2 Задаци

Задатак 7.1. Популација од 12 јединки подељена је на три кластера. Вредности обележја по кластерима су:

- 1. кластер: $2, 4, 6$;
- 2. кластер: $3, 3, 4, 5, 5$;
- 3. кластер: $1, 1, 2, 4$.

Методом простог случајног узорковања без понављања у узорак су одабрани први и трећи кластер. Наћи непристрасну оцену суме обележја популације и дисперзију те оцене.

Решење.

```
M <- 12 # veličina populacije
N <- 3 # broj klastera
n <- 2 # broj klastera koji ulaze u uzorak

# vrednosti obeležja jedinki po klasterima:
kl1 <- c(2, 4, 6)
kl2 <- c(3, 3, 4, 5, 5)
kl3 <- c(1, 1, 2, 4)
t_l <- c(sum(kl1), sum(kl2), sum(kl3)) # obeležja klastera
t_ocena <- N * mean(t_l[c(1, 3)])
t_ocena
```

```
## [1] 30
```

```
D_t_ocena <- N ^ 2 * (1 - n / N) * var(t_l) / n
D_t_ocena
```

```
## [1] 56
```

Задатак 7.2. Популација која садржи 100 јединки издељена је на 10 кластера. Изабран је прост случајан узорак од 3 кластера чије су суме обележја: 4, 12 и 7.
(а) Наћи непристрасну оцену укупне суме обележја популације и непристрасну оцену популацијске средине.
(б) Наћи оцене дисперзија оцена из (а).

Решење.

```
M <- 100 # veličina populacije
N <- 10 # broj klastera
n <- 3 # broj klastera koji ulaze u uzorak
t_l_na_uzorku <- c(4, 12, 7) # obeležja klastera

# a)
t_ocena <- N * mean(t_l_na_uzorku)
t_ocena
```

```
## [1] 76.66667
```

```
m_Y_ocena <- t_ocena / M
m_Y_ocena
```

```
## [1] 0.7666667
```

```
# b)
ocena_D_t_ocena <- N ^ 2 * (1 - n / N) * var(t_l_na_uzorku) / n
ocena_D_t_ocena
```

```
## [1] 381.1111
```

```
ocena_D_m_Y_ocena <- ocena_D_t_ocena / M ^ 2
ocena_D_m_Y_ocena
```

```
## [1] 0.03811111
```

Задатак 7.3. Истраживачи желе да процене укупан број оболелих од неке ретке болести у Србији. Посматрани су Борски, Рашки и Моравички округ и забележен је број оболелих по сваком округу.
Борски округ: 1001;
Рашки округ: 2230;
Моравички округ: 1798.
Укупан број округа је 29, а укупан број становника 7 186 862. Оценити укупан број оболелих у целој земљи и оценити дисперзију те оцене ако су посматрани окрузи изабрани методом простог случајног узорковања без понављања.

Решење.

```
M <- 7186862 # ukupan broj stanovnika
N <- 29 # broj klastera
n <- 3 # broj klastera koji ulaze u uzorak

t_l_na_uzorku <- c(1001, 2230, 1798)
```

```r
# Obeležje klastera je broj obolelih stanovnika u njima,
# jer je broj obolelih jednak sumi obeležja Y koje obolelom dodeljuje 1,
# a onom ko nije oboleo dodeljuje 0
t_ocena <- N * mean(t_l_na_uzorku)
t_ocena
```

```
## [1] 48613.67
```

```r
ocena_D_t_ocena <- N ^ 2 * (1 - n / N) * var(t_l_na_uzorku) / n
ocena_D_t_ocena
```

```
## [1] 97696366
```

Задатак 7.4. Популација која садржи 250 елемената издељена је на 20 кластера. Изабран је узорак од 5 различитих кластера са вероватноћама пропорционалним величинама кластера, чије су величине и суме обележја
M_l: $5, 17, 10, 12, 22$;
τ_l: $4, 11, 7, 11, 23$.
Наћи Horvitz-Thompson-ову оцену средње вредности обележја и оценити дисперзију те оцене.

Решење.

```r
M <- 250 # obim populacije
N <- 20 # broj klastera
n <- 5 # broj klastera koji su izabrani u uzorak

M_l <- c(5, 17, 10, 12, 22) # veličine klastera koji su izabrani u uzorak
t_l_na_uzorku <- c(4, 11, 7, 11, 23)

psi <- M_l / M # verovatnoće izbora klastera
pi <- 1 - (1 - psi) ^ n # verovatnoće uključenja klastera u uzorak klastera
m_Y_ht_ocena <- sum(t_l_na_uzorku / pi) / M
# svi klasteri koji su ušli u uzorak su različiti, pa je S' = S
m_Y_ht_ocena
```

```
## [1] 0.9175012
```

```r
ocena_D_m_Y_ht_ocena <- sum((1 - pi) * t_l_na_uzorku ^ 2 / (pi ^ 2))
for (i in 1:n) {
  for (j in 1:n) {
    if (i != j) {
      pi_ij <- pi[i] + pi[j] - 1 + (1 - psi[i] - psi[j]) ^ n
      ocena_D_m_Y_ht_ocena <-
        ocena_D_m_Y_ht_ocena +
        (pi_ij - pi[i] * pi[j]) * (t_l_na_uzorku[i] * t_l_na_uzorku[j]) /
        (pi[i] * pi[j] * pi_ij)
    }
  }
}
ocena_D_m_Y_ht_ocena <- ocena_D_m_Y_ht_ocena / (M ^ 2)
ocena_D_m_Y_ht_ocena
```

```
## [1] 0.005581875
```

Задатак 7.5. Дати су подаци о оценама из математике за 12 ученика који су подељени у 3 групе:
1. група: $2, 4, 5$;
2. група: $3, 3, 4, 5, 2$;
3. група: $1, 3, 5, 1$.
Бирањем пропорционалним величини групе, са понављањем, у узорак су одабране прва и трећа група. Наћи Hansen-Hurwitz-ову оцену просечне оцене из математике и дисперзију те оцене.

Решење.

```
M <- 12 # veličina populacije
N <- 3 # broj klastera
n <- 2 # broj klastera koji ulaze u uzorak
kl1 <- c(2, 4, 5)
kl2 <- c(3, 3, 4, 5, 2)
kl3 <- c(1, 3, 5, 1)
t_l <- c(sum(kl1), sum(kl2), sum(kl3)) # obeležje klastera
M_l <- c(length(kl1), length(kl2), length(kl3)) # veličine klastera

m_Y_ocena_hh <- mean(t_l[c(1, 3)] / M_l[c(1, 3)])
m_Y_ocena_hh
```

```
## [1] 3.083333
```

```
D_m_Y_ocena_hh <- sum(M_l * (t_l / M_l - sum(t_l) / M) ^ 2) / n / M
D_m_Y_ocena_hh
```

```
## [1] 0.1166667
```

Вежбе 8

Вишеетапни узорак. Систематски узорак

8.1 Неопходно предзнање

8.1.1 Вишеетапни узорак

Подсетимо се, кластер узорак смо добијали тако што од N примарних јединки одаберемо њих n, па затим из сваке од одабраних примарних, у узорак узмемо **све** секундарне јединке. Овакав начин узорковања може бити веома неисплатив уколико су обими кластера велики, а уколико у кластерима постоји много међусобно сличних секундарних јединки. Улазак свих таквих секундарних јединки неће нимало (или скоро нимало) повећати репрезентативност узорка, па је овај поступак неопходно „поправити''.

Праволинијско решење за овај проблем јесте то да се из сваког одабраног кластера не задржавају све секундарне јединке, већ да се из тог кластера извуче узорак, и да те секундарне јединке уђу у нови, тзв. **двоетапни** узорак. Класични кластер узорак се, из јасних разлога, стога често назива и једноетапни узорак.

Оцене параметара почетне популације ће, наравно, варирати у зависности од тога по ком плану је вршено узорковање примарних и секундарних јединки. Овде ћемо се држати ознака које смо користили на претходним вежбама, код (једноетапног) кластер узорковања. Подсетимо их се:

- y_{lk} - вредност обележја Y на k-тој јединки из l-тог кластера;
- N - укупан број кластера;
- n - број кластера одабраних у узорак;
- M_l - величина l-тог кластера;
- $M = \sum_{l=1}^{N} M_l$ - обим популације;
- τ_l - сума обележја на l-том кластеру;
- $\tau_Y = \sum_{l=1}^{N} \tau_l$ - сума обележја на целој популацији;
- $m_l = \frac{\tau_l}{M_l}$ - средина l-тог кластера;
- $\sigma_l^2 = \frac{1}{M_l - 1} \sum_{k=1}^{M_l} (y_{lk} - m_l)^2$ - дисперзија l-тог кластера.

Са n_l ћемо означити број секундарних јединки које се бирају из l-тог кластера. Ми ћемо изложити оцене које се добију када се обе етапе узорковања изводе по плану ПСУ без понављања. Наравно, то је само једна од могућности, али се остале оцене добијају у аналогији са овима, што се по потреби у реалном истраживању може пронаћи у литератури.

Двоетапни узорак код кога се примарне и секундарне јединице бирају као ПСУ без понављања

За оцену тотала τ_Y предлаже се

$$\hat{\tau}_Y^u = \frac{N}{n} \sum_{l \in S} \hat{\tau}_l,$$

где је $\hat{\tau}_l$ оцена тотала l-тог кластера, која се на основу ПСУ без понављања добија као

$$\hat{\tau}_l = \frac{M_l}{n_l} \sum_{k \in S_l} y_{lk},$$

где је S узорак кластера, а S_l узорак јединки из l-тог кластера из S.

Ова оцена је непристрасна, тј. важи $\mathbf{E}\hat{\tau}_Y^u = \tau_Y$, а њена дисперзија је:

$$\mathbf{D}\hat{\tau}_Y^u = \frac{N^2}{n}\left(1 - \frac{n}{N}\right)\sigma_\tau^2 + \frac{N}{n}\sum_{l=1}^{N} \frac{M_l^2}{n_l}\left(1 - \frac{n_l}{M_l}\right)\sigma_l^2,$$

где је σ_τ^2 дисперзија популације кластера, односно $\sigma_\tau^2 = \frac{1}{N-1}\sum_{l=1}^{N}\left(\tau_l - \frac{\tau_Y}{N}\right)^2$.
Непристрасна оцена ове дисперзије је:

$$\hat{\mathbf{D}}\hat{\tau}_Y^u = \frac{N^2}{n}\left(1 - \frac{n}{N}\right)\overline{S}_\tau^2 + \frac{N}{n}\sum_{l \in S} \frac{M_l^2}{n_l}\left(1 - \frac{n_l}{M_l}\right)\overline{S}_l^2,$$

где су

$$\overline{S}_\tau^2 = \frac{1}{n-1}\sum_{l \in S}\left(\hat{\tau}_l - \frac{\hat{\tau}_Y^u}{N}\right)^2 \quad \text{и} \quad \overline{S}_l^2 = \frac{1}{n_l - 1}\sum_{k \in S_l}\left(y_{lk} - \frac{1}{n_l}\sum_{k \in S_l} y_{lk}\right)^2.$$

За оцену параметра m_Y и њене особине важи:

$$\hat{m}_Y^u = \frac{\hat{\tau}_Y^u}{M}, \;\; \mathbf{E}\hat{m}_Y^u = m_Y.$$

$$\mathbf{D}\hat{m}_Y^u = \frac{\mathbf{D}\hat{\tau}_Y^u}{M^2}, \;\; \hat{\mathbf{D}}\hat{m}_Y^u = \frac{\hat{\mathbf{D}}\hat{\tau}_Y^u}{M^2}.$$

Количничке оцене

Уместо горе поменутих оцена, можемо користити и количничку оцену, где ћемо као помоћно обележје посматрати величине кластера (уколико то има смисла):

$$\hat{\tau}_Y^r = b \cdot M,$$

где је

$$b = \frac{\sum_{l\in S} \hat{\tau}_l}{\sum_{l\in S} M_l}.$$

Дисперзија ове оцене је приближно једнака:

$$\mathbf{D}\hat{\tau}_Y^r \approx \frac{N^2}{n}\left(1-\frac{n}{N}\right)\frac{1}{N-1}\sum_{l=1}^{N}(\tau_l - BM_l)^2 + \frac{N}{n}\sum_{l=1}^{N}\frac{M_l^2}{n_l}\left(1-\frac{n_l}{M_l}\right)\sigma_l^2,$$

где је

$$B = \frac{\sum_{l=1}^{N}\tau_l}{\sum_{l=1}^{N} M_l} = \frac{\tau_Y}{M} = m_Y.$$

За оцену дисперзије предлаже се:

$$\hat{\mathbf{D}}\hat{\tau}_Y^r = \frac{N^2}{n}\left(1-\frac{n}{N}\right)\frac{1}{n-1}\sum_{l\in S}(\hat{\tau}_l - bM_l)^2 + \frac{N}{n}\sum_{l\in S}\frac{M_l^2}{n_l}\left(1-\frac{n_l}{M_l}\right)\overline{S}_l^2.$$

За оцену параметра m_Y и њене особине важи:

$$\hat{m}_Y^r = \frac{\hat{\tau}_Y^r}{M} = b,$$

$$\mathbf{D}\hat{m}_Y^r = \frac{\mathbf{D}\hat{\tau}_Y^r}{M^2},\ \ \hat{\mathbf{D}}\hat{m}_Y^r = \frac{\hat{\mathbf{D}}\hat{\tau}_Y^r}{M^2}.$$

8.1.2 Систематски узорак

Систематски узорак, као што ћемо ускоро видети, представља начин узорковања који је веома једноставан за спровођење, али чија прецизност може драстично да варира у зависности од распореда јединки у популацији.

За почетак, вратимо се старим ознакама. Нека нам је N обим популације, а n обим узорка. Нека су за почетак N и n такви да је $k = \frac{N}{n}$ природан број. Узорак се добија тако што се на случајан начин одабере број од 1 до k, а потом у узорак улази свака k-та јединка. Такав узорак називамо **систематски узорак**. Уколико $\frac{N}{n}$ није цео број, посебно се дефинише како се систематски узорак формира, углавном „са прекорачењем''.

Оцена популацијске средине на основу систематског узорка и њене особине су:

$$\hat{m}_Y^{sys} = \frac{\sum_{i\in S} y_i}{n}, \quad \mathbf{E}\hat{m}_Y^{sys} = m_Y.$$

$$\mathbf{D}\hat{m}_Y^{sys} = \frac{N-1}{N}\sigma_Y^2 - \frac{k(n-1)}{N}\sigma_{sys}^2,$$

где је

$$\sigma_{sys}^2 = \frac{1}{k(n-1)}\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \overline{Y}_i)^2,$$

а y_{ij} је j-ти елемент i-тог могућег узорка, а $\overline{Y}_i$ његова узорачка средина.
За оцену тотала и њене особине важи:

$$\hat{\tau}_Y^{sys} = N\hat{m}_Y^{sys}, \quad \mathbf{E}\hat{\tau}_Y^{sys} = \tau_Y, \quad \mathbf{D}\hat{\tau}_Y^{sys} = N^2\mathbf{D}\hat{m}_Y^{sys}.$$

8.2 Задаци

Задатак 8.1. У датотеци `skupina.txt` налазе се вредности обележја за популацију од 500 јединки, која је подељена на 3 кластера. Бира се прост случајан узорак од две примарне јединице, а затим се из сваке одабране примарне јединице бира по 10 секундарних јединица. Оценити суму обележја на популацији и одредити оцену дисперзије те оцене.

Решење.

```
skupina <- read.table("skupina.txt")
head(skupina)
```

```
##   kl yi
## 1  1 77
## 2  3 81
## 3  3 60
## 4  2 89
## 5  2 25
## 6  2 91
```

```
M <- length(skupina$kl) # obim populacije
N <- 3 # broj klastera
n <- 2 # broj klastera koji ulaze u uzorak

klasteri <- list()
Ml <- c()
tau_l <- c()
for(i in 1:3) {
  klasteri[[i]] <- skupina[skupina$kl==i, ]
  Ml[i] <- length(klasteri[[i]]$kl)
  tau_l[i] <- sum(klasteri[[i]]$yi)
}
```

```r
indeks_kl <- sample(1:3, 2)
# klasteri odbrani u uzorak
klasteri_uz <- list(klasteri[[indeks_kl[1]]], klasteri[[indeks_kl[2]]])
Ml_kl <- Ml[indeks_kl] # veličine klastera odabranih u uzorak

# iz svakog klastera odabranog u uzorak biramo 10 jedinki
uz <- list()
for(i in 1:length(klasteri_uz)) {
  uz[[i]] <- sample(klasteri_uz[[i]]$yi, 10)
}
uz # vrednosti obeležja za odabrane sekundarne jedinice
```

```
## [[1]]
##  [1] 37 66 40 62 95 63 27 49 85 46
##
## [[2]]
##  [1] 88  4 94 49  8 16 92 19 84 56
```

```r
# ocene na uzorcima
tau_l_ocene <- c()
for(i in 1:length(uz)) {
  tau_l_ocene[i] <- Ml_kl[i] * mean(uz[[i]])
}

t_ocena <- N * sum(tau_l_ocene) / n
t_ocena
```

```
## [1] 26739
```

```r
# ocena disperzije za t_ocena
S_t2 <- sum((tau_l_ocene - t_ocena / N) ^ 2) / (n - 1)

nl <- c(10, 10)
S_l2 <- c()
for (i in 1:length(uz)) {
  S_l2[i] <- var(uz[[i]])

}

ocena_D_t_ocena <-
  N ^ 2 * (1 - n / N) * S_t2 / n + N * sum(Ml_kl ^ 2 * (1 - nl / Ml_kl) * S_l2 / nl) / n
ocena_D_t_ocena
```

```
## [1] 7880336
```

Задатак 8.2. У околини једног града налази се 7 села, која представљају кластере. Циљ истраживања је да се одреди просечан број чланова домаћинства. Изабран је прост случајан узорак од 3 села: прво, треће и шесто. Из сваког одабраног села одабрано је неколико домаћинстава и забележени су бројеви чланова. Добијени су следећи резултати:
Прво село: $2, 5, 4, 2, 3, 5, 3, 1$;
Треће село: $4, 5, 5, 6, 1, 2, 2, 3, 3, 4$;
Шесто село: $7, 6, 6, 4, 2, 1, 3$.
Број домаћинстава у седам села је, редом, $190, 340, 450, 25, 78, 80, 107$. Оценити просечан број чланова домаћинства количничком оценом.

Решење.

```
N <- 7 # broj sela
n <- 3 # u uzorak je odabrano prvo, treće i šesto selo
uz1 <- c(2, 5, 4, 2, 3, 5, 3, 1) # broj članova domaćinstava odabranih iz prvog sela
uz2 <- c(4, 5, 5, 6, 1, 2, 2, 3, 3, 4) # broj članova domaćinstava
                                       # odabranih iz trećeg sela
uz3 <- c(7, 6, 6, 4, 2, 1, 3) # broj članova domaćinstava odabranih iz šestog sela
uz <- list(uz1, uz2, uz3)

Ml <- c(190, 340, 450, 25, 78, 80, 107) # broj domaćinstava u svih sedam sela
M <- sum(Ml) # ukupno domaćinstava u svih sedam sela
Ml_uz <- c(Ml[1], Ml[3], Ml[6]) # broj domaćinstava u selima odabranim u uzorak

nl <- as.numeric(lapply(uz, length)) # broj odabranih domaćinstava iz svakog sela
                                     # odabranog u uzorak
# lapply(uz, length) - na svaki element liste "uz" primenjuje "length" - vraca listu,
# pa as.numeric - da pretvorimo u vektor
nl
```

```
## [1]  8 10  7
```

```
tau_l_ocene <- c()
for (i in 1:length(uz)) {
  tau_l_ocene[i] <- Ml_uz[i] * mean(uz[[i]])
}

b <- sum(tau_l_ocene) / sum(Ml_uz) # videli smo da je b jednako količničkoj
                                   # oceni populacijske sredine
b
```

```
## [1] 3.47247
```

Задатак 8.3. Популација се састоји од 12 јединки, а обележје је редни број јединке. Формирати систематске узорке обима 4, а затим одредити дисперзију средине систематског узорка.

Решење.

```
N <- 12
n <- 4

# korak
```

```r
k <- N / n
k # jeste ceo broj
```

```
## [1] 3
```

```r
# mogući uzorci obima n su dati u kolonama (ima ih k)
uzorci <- matrix(1:12, n, k, byrow = T)
uzorci
```

```
##      [,1] [,2] [,3]
## [1,]    1    2    3
## [2,]    4    5    6
## [3,]    7    8    9
## [4,]   10   11   12
```

```r
# disperzija
sigma_2 <- var(as.numeric(uzorci)) # as.numeric(uzorci) - matricu pretvori u vektor
sigma_sis_2 <-  0
for(i in 1:k) {
  for (j in 1:n) {
    sigma_sis_2 <- sigma_sis_2 + (uzorci[j, i] - mean(uzorci[, i])) ^ 2
  }
}

sigma_sis_2 <- sigma_sis_2 / (k * (n - 1))

D_m_sis <- (N - 1) * sigma_2 / N - k * (n - 1) * sigma_sis_2 / N
D_m_sis
```

```
## [1] 0.6666667
```

Задатак 8.4. У датотеци `stanovi.txt` налазе се подаци о ценама 500 станова у 4 београдске општине. У првој колони дата је општина, а у другој цена стана.

(а) Нека општине представљају кластере. Извадити узорак од 2 кластера методом простог случајног узорковања без понављања и оценити просечну цену станова, а затим одредити дисперзију те оцене.

(б) Извадити систематски узорак обима 100, оценити просечну цену станова и одредити дисперзију те оцене.

Решење.

```r
stanovi <- read.table("stanovi.txt")
head(stanovi)
```

```
##    opstina  cena
## 1 Vozdovac 59469
## 2 Cukarica 57569
## 3 Cukarica 79374
## 4 Vozdovac 60413
## 5 Cukarica 89394
## 6   Vracar 54382
```

```r
# a)
# formiramo klastere
```

```r
kl <- list()
kl[[1]] <- stanovi[stanovi$opstina == "Vracar",]
kl[[2]] <- stanovi[stanovi$opstina == "Vozdovac",]
kl[[3]] <- stanovi[stanovi$opstina == "Cukarica",]
kl[[4]] <- stanovi[stanovi$opstina == "Rakovica",]

N <- 4 # 4 klastera
n <- 2 # 2 klastera ulaze u uzorak
Ml <- c() # veličine klastera
tau_l <- c() # obeležja klastera
for(i in 1:length(kl)) {
  Ml[i] <- length(kl[[i]]$opstina)
  tau_l[i] <- sum(kl[[i]]$cena)
}
M <- sum(Ml)

indeks <- sample(1:4, 2) # biramo dva klastera

m_Y_ocena <- N * mean(tau_l[indeks]) / M
m_Y_ocena
```

```
## [1] 86509.87
```

```r
D_m_Y_ocena <- N ^ 2 * (1 - n / N) * var(tau_l) / n / M ^ 2
D_m_Y_ocena
```

```
## [1] 1538653
```

```r
# b)

N <- M # obim populacije
n <- 100 # obim sistematskog uzorka
k <- N / n

sis_uzorci <- matrix(stanovi$cena, n, k, byrow = TRUE) # mogući uzorci su dati u kolonama
sis_uzorak <- sis_uzorci[, sample(1:ncol(sis_uzorci), 1)] # na slučajan način biramo
                                                          # jednu kolonu - uzorak
sis_uzorak
```

```
##   [1]  59469  54382  82999  54514  81418  74751  66124  89491  68717  54524
##  [11]  71764  52719  69647  66582  83328  82906  53436  87797  50292  88530
##  [21]  81637  77595  55840  80127  71192  88830  84370  71640  50656  73641
##  [31]  76603  75297  58530  66072  80616  67059  52642  63232  77178  66118
##  [41]  66082  56285  85662  55906  87585  87291  57695  73907  51695  72704
##  [51]  72461  88802  77700  82851  59383  62958  85645  59948  89803  77300
##  [61]  87796  64743  51119  66112  51903  73675  56887  60987  63186  60586
##  [71]  83068  57591  50814  57400  84241  62931  57350  60810  65670  54570
##  [81]  52966  79926  78814  76376  76490  81406  72486  74260  78693  71678
##  [91]  89260  84612 222596 139816 166313 425139 138098 209319 463466 475322
```

```r
m_sis <- mean(sis_uzorak)
m_sis # ocena prosecne cene
```

```
## [1] 86924.03
```

```r
# disperzija
sigma_2 <- var(as.numeric(sis_uzorci))
sigma_sis_2 <- 0
for(i in 1:k) {
  for (j in 1:n) {
    sigma_sis_2 <- sigma_sis_2 + (sis_uzorci[j, i] - mean(sis_uzorci[, i])) ^ 2
  }
}

sigma_sis_2 <- sigma_sis_2 / (k * (n - 1))

D_m_sis <- (N - 1) * sigma_2/N - k * (n - 1) * sigma_sis_2 / N
D_m_sis
```

```
## [1] 8281740
```

Вежбе 9

Непараметарска статистика. Џекнајф

Напомена. Овај део градива, одавде па до краја скрипте, је информативног карактера и не долази у обзир за писмени испит.

Напомена. О овоме сам први пут чуо на курсу Одабрана поглавља математичке статистике код др Марка обрадовића и др Марије Цупарић. Послужио сам се њиховим концептима и идејама на много места, а додао сам и неке ствари из разне литературе. Свакако, не сматрам се идејним творцем оваквог концепта излагања ове материје, те не полажем никаква ауторска права.

9.1 Непараметарска статистика

Од сад па на даље бавићемо се простим случајним узорком **са понављањем** и то са становишта **приступа модела**. Другим речима, наш узорак биће коначан низ случајних величина $X_1, \dots, X_n$ где су све међусобно независне и једнако расподељене.

Наиме, до сада смо се углавном бавили двама оценама: оценом средње вредности и оценом дисперзије популације. Када то „пребацимо'' на приступ модела, ми имамо узорак из неке расподеле вероватноћа са функцијом расподеле $F(x;\theta)$, где θ представља један или више параметара расподеле. Рецимо, уколико говоримо о нормалној расподели, $\theta = (\mu, \sigma^2)$, где је μ очекивање расподеле, а σ^2 њена дисперзија. Код експоненцијалне расподеле $\mathcal{E}(\lambda)$, параметар је λ, и он није једнак ни очекивању ни дисперзији расподеле, већ је просто неки параметар од којег она зависи.

Све ово поменуто проучава грана статистике која се често назива „параметарска статистика''. Суштински, позната нам је класа расподела, „до на параметар'': знамо да је узорак из нормалне расподеле, али не знамо колико је μ, знамо да је из експоненцијалне, али не знамо колико је λ итд. За оцене оваквих параметара, као и особина тих оцена, попут дисперзије, стандардне девијације, интервала поверења итд. развијене су разни методи, попут метода замене, метода максималне веродостојности и слично.

Многи параметри θ, пак, зависе директно од функције расподеле, то јест они су **функција функције расподеле**, и често их зовемо **статистичким функционалима**. Рецимо, очекивање је један такав параметар:

$$\theta = \mathbf{E}X = \int x \mathrm{d}F(x).$$

И дисперзија такође:

$$\mathbf{D}X = \mathbf{E}(X - \mathbf{E}X)^2 = \int \left[x - \int x \mathrm{d}F(x)\right]^2 \mathrm{d}F(x).$$

Јасно, овде спадају и моменти, центрирани моменти, стандардна девијација итд. Грана статистике која, између осталог, изучава статистичке функционале, често се назива **непараметарска статистика**. Сходно свему реченом, ваљало би да постоје неки **универзални методи** који испитују квалитет оцене параметра/статистичког функционала. И постоје.

9.2 Подузорковање (енг. *Resampling*)

Постоји неколико метода за оцењивање параметара које су засноване на подузорковању, а најпознатије су џекнајф и бутстреп. Један ћемо радити на овим вежбама, а други на следећим.

Методе засноване на подузорковању темеље се на томе да се узорак који имамо накратко посматра као популација, те да се из њега вуку узорци, мањег обима од његовог обима. Основна нада лежи у томе да се оваквим приступом може „убити'' **нерепрезентативност** узорка, то јест појава да се у узорку нађу вредности за које је мало вероватно да се ту нађу. Дакле, из узорка се ваде „подузорци'', на сваком од њих се рачуна вредност параметра од интереса, и те вредности се пореде тражећи аномалије. Ако има драстичних варијација међу тим вредностима, узорак је вероватно нерепрезентативан.

Наравно, ово је само мотивација за сам приступ, а његове остале предности и мане демонстрираћемо конкретније и „математичкије'' на примерима џекнајфа и бутстрепа.

9.3 Џекнајф метод

Нека имамо оцену T_n (неког параметра θ) засновану на узорку $X_1, \ldots, X_n$. Џекнајф метод знаснива се на рачунању оцене T_{n-1} за све подузорке тог реда (тј. сваки пут склонимо по један елемент из узорка и направимо оцену).

Претпоставимо да је $\mathbf{E}T_n = \theta + \mathbf{b}(T_n)$. Обележимо са $T_{n,-i}$ оцену на основу узорка из ког је избачен елемент X_i и нека је $T_{n,\bullet}$ њихова средња вредност, тј.

$$T_{n,\bullet} = \frac{1}{n}\sum_{i=1}^{n} T_{n,-i}.$$

Џекнајф оцена пристрасности се тада дефинише као

$$\hat{\mathbf{b}}_{\mathbf{jack}}(T_n) = (n-1)(T_{n,\bullet} - T_n).$$

После овако оцењене пристрасности коригована оцена је

$$T_{n,\mathbf{jack}} = T_n - \hat{\mathbf{b}}_{\mathbf{jack}}(T_n).$$

Одакле ова идеја? За многе статистике може се показати да је, за неке a и b,

$$\mathbf{b}(T_n) = \frac{a}{n} + \frac{b}{n^2} + O\left(\frac{1}{n^3}\right). \tag{9.1}$$

Узмимо прво случај да је $\mathbf{b}(T_n) = a/n$ без чланова нижег реда величине (то је нпр. случај оцене дисперзије). Тада је

$$\mathbf{E}T_{n,\bullet} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{E}T_{n,-i} = \theta + \frac{a}{n-1},$$

па је

$$\mathbf{E}T_{n,\mathbf{jack}} = \mathbf{E}(T_n - \hat{\mathbf{b}}_{\mathbf{jack}}(T_n)) = \theta + \frac{a}{n} - (n-1)\left(\frac{a}{n-1} - \frac{a}{n}\right) = \theta.$$

Видимо да је у овом случају $\hat{\mathbf{b}}_{\mathbf{jack}}(T_n)$ непристрасна оцена $\mathbf{b}(T_n)$. У општем случају 9.1 имамо да је

$$\mathbf{b}(T_{n,-i}) = \frac{a}{n-1} + \frac{b}{(n-1)^2} + O(1/n^3),$$

а исти облик наравно има и $\mathbf{b}(T_{n,\bullet})$. Добијамо

$$\begin{aligned}
\mathbf{E}\hat{\mathbf{b}}_{\mathbf{jack}}(T_n) &= (n-1)(\mathbf{E}T_{n,\bullet} - \mathbf{E}T_n) \\
&= (n-1)\left[\left(\frac{a}{n-1} - \frac{a}{n}\right) + \left(\frac{b}{(n-1)^2} - \frac{b}{n^2}\right) + O\left(\frac{1}{n^3}\right)\right] \\
&= \frac{a}{n} + \frac{(2n-1)b}{n^2(n-1)} + O\left(\frac{1}{n^2}\right) \\
&= \mathbf{b}(T_n) + O\left(\frac{1}{n^2}\right).
\end{aligned}$$

У општем случају оцена није непристрасна, али је пристрасност смањена за ред величине у односу на полазну оцену.

Пример 9.1. Имамо расподелу чији је десни крај носача θ коначан и непознат. У том случају разумна оцена тог параметра је $X_{(n)}$. Ова оцена ће увек бити пристрасна и њена средња вредност ће бити мања од θ.
Типични пример је униформна расподела на интервалу $[0, \theta]$. У том случају је средња вредност оцене

$$\mathbf{E}X_{(n)} = \frac{n}{n+1}\theta = \theta\left(1 - \frac{1}{n} + \frac{1}{n^2} - \frac{1}{n^3} + \cdots\right),$$

па џекнајф оцена има смисла. Приметимо да је[a]

$$X_{(n),-i} = \begin{cases} X_{(n)}, & i < n \\ X_{(n-1)}, & i = n, \end{cases}$$

па добијамо да је

$$X_{(n),\bullet} = \frac{n-1}{n}X_{(n)} + \frac{1}{n}X_{(n-1)},$$

а џекнајф оцена

$$X_{(n),\mathbf{jack}} = X_{(n)} + \frac{n-1}{n}(X_{(n)} - X(n-1)).$$

Пристрасност ове оцене је реда $1/n^2$ и свакако је мања од пристрасности полазне оцене. Наравно, оцену смо могли веома једноставно начинити и непристрасном, множећи је са $(n+1)/n$. То је једна од мана џекнајфа и сличних општих метода: у конкретном случају, скоро увек постоји бољи метод. Међутим, ако имамо непараметарски проблем с почетка примера, џекнајф оцена ће и даље смањивати пристрасност и, као таква, може нам бити од користи, док оцена $\frac{n+1}{n}X_{(n)}$, која је била најбоља за случај униформне расподеле, у општем случају не само да неће бити непристрасна, већ се лако може догодити да има много већу пристрасност од полазне.

[a]Мало смо злоупотребили ознаке. Под $X_{(n)}$ подразумевамо највећи елемент узорка, иако их овде има $n-1$ након избацивања.

Пошто смо оценили пристрасност, позабавићемо се дисперзијом, као још једном честом мером квалитета оцене. Џон Тјуки је предложио аналогни метод за њено оцењивање. Његова џекнајф оцена дисперзије је

$$\widehat{\mathbf{D}}_{\mathbf{jack}}(T_n) = \frac{n-1}{n}\sum_{i=1}^{n}(T_{n,-i} - T_{n,\bullet})^2. \tag{9.2}$$

За разлику од оцене пристрасности, оцена за дисперзију није баш интуитивна. Откуд ова идеја? Пређутно се претпоставља да се оцена може приближно изразити као средина независних величина (што јесте случај с великом класом оцена),

$$T_n \approx \frac{1}{n}\sum_{i=1}^{n}\Psi(X_i).$$

Тада би дисперзија оцене била

$$\mathbf{D}T_n \approx \frac{1}{n}\mathbf{D}\Psi(X_i).$$

Приликом извођења Тјуки је пошао од кориговане оцене $T_{n,\mathbf{jack}}$ која се може представити и као

$$T_{n,\mathbf{jack}} = \frac{1}{n}\sum_{i=1}^{n}(nT_n - (n-1)T_{n,-i}) = \frac{1}{n}\sum_{i=1}^{n}\tilde{T}_i. \tag{9.3}$$

Сабирке $\tilde{T}_i$ (који представљају $\Psi(X_i)$) назвао је *псеудовредностима*, а $T_{n,\mathbf{jack}}$ је њихова „узорачка'' средина. Њихова поправљена узорачка дисперзија тада је

$$\begin{aligned}\frac{1}{n-1}\sum_{i=1}^{n}(\tilde{T}_i - \overline{\tilde{T}}_i)^2 &= \frac{1}{n-1}\sum_{i=1}^{n}\Big((n-1)(T_{n,\bullet} - T_{n,-i})\Big)^2 \\ &= (n-1)\sum_{i=1}^{n}(T_{n,-i} - T_{n,\bullet})^2.\end{aligned} \tag{9.4}$$

Тјукијева џекнајф оцена дисперзије сада се добија када се 9.4 подели са n.

Под неким условима џекнајф оцена дисперзије је постојана, тј. важи

$$\frac{\widehat{\mathbf{D}}_{\mathbf{jack}} T_n}{\mathbf{D} T_n} \xrightarrow{P} 1.$$

На пример, у случају када је $T_n = g(\overline{X}_n)$, а g непрекидно диференцијабилна функција. Међутим, у неким случајевима то не важи. Ефрон је показао да џекнајф оцена квантила није постојана, штавише, у случају медијане показао је да

$$\frac{\widehat{\mathbf{D}}_{\mathbf{jack}} M_n}{\mathbf{D} M_n} \xrightarrow{d} Y^2, \tag{9.5}$$

где Y има $\mathcal{E}(1)$ расподелу.

Предложена су и побољшања џекнајф процедуре. Једна од најпознатијих је џекнајф с брисањем d података, тј. уместо подузорака обима $n-1$ радимо с подузорцима обима $n-d$. Тих узорака има $\binom{n}{d}$. Поређајмо те подузорке у низ и нека је $T_{n,d,\alpha}$, $1 \leq \alpha \leq \binom{n}{d}$ оцена на основу подузорка редног броја α, а $s^2_{n,d}$ узорачка дисперзија вредности $T_{n,d,\alpha}$, $1 \leq \alpha \leq \binom{n}{d}$. Џекнајф оцена дисперзије овом процедуром тада је

$$\widehat{\mathbf{D}}_{\mathbf{jack},d} = \frac{n-d}{d} s^2_{n,d}.$$

Показује се да се најбољи резултати добијају ако $d \to +\infty$ када $n \to +\infty$. Стога је јасно да је ова процедура рачунски презахтевна, па се у пракси она апроксимира тако што се уместо свих подузорака величине $n-d$ случајно бира одређен број m и на основу њега изводи џекнајф оцена дисперзије.

9.3.1 Рачунарска имплементација

Како је овај део курса информативан, нећемо пролазити кроз конкретне задатке, јер се имплементација разликује од језика до језика. Рецимо, у програмском језику `R`, у пакету `bootstrap`, постоји функција `jackknife` која има следећу синтаксу:

`jackknife(uzorak, funkcija_koja_racuna_parametar)`,

и она враћа листу која садржи поправљену џекнајф оцену параметра, џекнајф оцену пристрасности, вредности оцене за сваки подузорак и на крају ниску са самим позивом функције. Слично је и у осталим програмским језицима.

Вежбе 10

Бутстреп

Видели смо да је џекнајф метод згодан метод за оцењивање квалитета оцена. Такође, видели смо да он ради добро у неким случајевима, док у другим и није баш препоручљив за коришћење. Ефрон је у једном свом раду из 1979. године водећи се сличном идејом дао основе метода који нам у много више случајева може помоћи да лакше оценимо одговарајуће параметре.

Нека нам је од интереса статистички функционал (параметар) $h(F)$ који смо оценили **методом замене**: оцена за $h(F)$ је $h(\hat{F}_n)$, где је $\hat{F}_n$ емпиријска функција расподеле.

Неке од мера квалитета су

а) Расподела грешке $\lambda_n(F) = \mathbf{P}_F(\sqrt{n}(h(\hat{F}_n) - h(F)) \leq a)$ или $\lambda_n(F) = \mathbf{P}_F\left(\sqrt{n}\frac{h(\hat{F}_n)-h(F)}{\gamma(F)} \leq a\right)$;

б) Пристрасност оцене $\lambda_n(F) = \mathbf{b}(h(\hat{F}_n))$;

в) Дисперзија оцене $\lambda_n(F) = \mathbf{D}(h(\hat{F}_n))$.

На пример, желимо да оценимо вероватноћу под а). Циљ нам је расподела случајне величине $\lambda_n(\hat{F}_n)$. Природно је покушати принципом замене и тада се оцена добија када се расподела F свих елемената узорка замени са $\hat{F}_n$. Тада $h(F)$ постаје $h(\hat{F}_n)$. Али чиме заменити $h(\hat{F}_n)$ која већ зависи од $\hat{F}_n$?

Оцена $h(\hat{F}_n)$ може се приказати као оцена $h(X_1, \dots, X_n)$, тј. не преко узорачке функције већ преко самог узорка, тј.

$$h(\hat{F}_n) = h(X_1, \dots, X_n).$$

Пошто је $\hat{F}_n$ оцена за F, оцена се изражава преко узорка из F. Да бисмо добили оцену за $h(\hat{F}_n)$, узимамо одговарајући узорак $X_1^*, \dots, X_n^*$ из $\hat{F}_n$, па је оцена за $h(\hat{F}_n)$

$$h_n^* = h(X_1^*, \dots, X_n^*).$$

Тада се $\lambda_n(\hat{F}_n)$ може записати као

$$\lambda_n(\hat{F}_n) = \mathbf{P}_{\hat{F}_n}\left(\sqrt{n}(h_n^* - h(\hat{F}_n)) \leq a\right).$$

Остаје питање шта значи да узмемо узорак из $\hat{F}_n$. Нама и не треба конкретан узорак (који немамо), већ нам је циљ одредити вероватноћу да хипотетички узорак из ове расподеле задовољи неки услов. Знамо да је расподела за $\hat{F}_n$ дискретна униформна на скупу реализованих вредности $x_1, \dots, x_n$, а $X_1^*, \dots, X_n^*$ је просто хипотетички узорак из ове униформне расподеле. Размотримо на примеру.

Пример 10.1 (оцењивање расподеле грешке оцене очекивања)**.** Нека имамо параметар

$$\theta = h(F) = \mathbf{E}_F(X).$$

Јасно, тада је $h(\hat{F}_n) = \overline{X}_n$ (рачунали смо негде раније). Рецимо да желимо да срачунамо

$$\lambda_n(\hat{F}_n) = \mathbf{P}_{\hat{F}_n}\left(\sqrt{n}(\theta_n^* - h(\hat{F}_n)) \le a\right) \tag{10.1}$$

у нереалистичном случају $n = 2$. Претпоставимо да су статистике поретка дате са

$$X_{(1)} = c < X_{(2)} = d.$$

Тада су X_1^* и X_2^* независне, са истом расподелом

$$\mathbf{P}(X_i = c) = \mathbf{P}(X_i = d) = \frac{1}{2},\ i = 1, 2.$$

Пар (X_1^*, X_2^*) стога узима један од могућа четири пара вредности

$$(c, c),\ (c, d),\ (d, c),\ (d, d),$$

сваки са вероватноћом $1/4$. Стога

$$\theta^* = \frac{1}{2}(X_1^* + X_2^*)$$

узима вредности c, $\frac{1}{2}(c + d)$ и d са вероватноћама $\frac{1}{4}$, $\frac{1}{2}$ и $\frac{1}{4}$ редом, одакле

$$\theta^* - h(\hat{F}_n) = \theta^* - \frac{1}{2}(c + d)$$

узима вредности $\frac{1}{2}(c - d)$, 0, $\frac{1}{2}(d - c)$ са вероватноћама $\frac{1}{4}$, $\frac{1}{2}$ и $\frac{1}{4}$ редом, одакле сада можемо рачунати 10.1 за било коју вредност a.

Пример 10.2 (бутстреповање вероватноће)**.** Нека желимо да оценимо:

$$\lambda_n(F) = \mathbf{P}_F\Big((X_1, \dots, X_n) \in A\Big).$$

Тада је $\lambda_n(\hat{F}_n)$ облика

$$\lambda_n(\hat{F}_n) = \mathbf{P}_{\hat{F}_n}\Big((X_1^*, \dots, X_n^*) \in A\Big),$$

где је $X_1^*, \dots, X_n^*$ узорак из $\hat{F}_n$, а A неки одређен скуп. Другим речима, имамо проблем да одредимо вероватноћу да случајни вектор (с независним и једнако расподељеним компонентама) упадне у неки одређен скуп, што се своди на n-троструку суму или интеграл.
За велико n ово нам ствара својеврсни проблем: ми теоријски знамо како рачунати бутстреп оцену, али је она рачунски захтевна/немогућа.

Стандардни приступ овом проблему је Монте Карло апроксимација. Симулирамо вредности вектора $(X_1^*, \dots, X_n^*)$ велики број (B) пута па ће релативна фреквенција (удео) оних реализација које су упале у скуп A бити добра апроксимација вероватноће. Ово знамо на основу Бернулијевог закона великих бројева. Значи треба да симулирамо B „узорака" из расподеле $\hat{F}_n$, тј.

$$X_{11}^*, \dots, X_{1n}^*$$
$$X_{21}^*, \dots, X_{2n}^*$$
$$\dots$$
$$X_{B1}^*, \dots, X_{Bn}^*.$$

Добијени узорци називају се бутстреп узорци. Монте Карло апроксимација за $\lambda_n(\hat{F}_n)$ биће

$$\lambda_{B,n}^* = \frac{1}{B}\sum_{i=1} I\{(X_{i1}^*, \dots, X_{in}^*) \in A\}.$$

Величина $\lambda_{B,n}^*$ назива се бутстреп апроксимацијом. Наравно, ова процедура није ограничена само на случај када оцењујемо непознату расподелу оцене (вероватноћу), већ је применљива и на остале функционале који зависе од n. Цела двофазна процедура, оцењивање $\lambda_n(F)$ помоћу $\lambda_n(\hat{F}_n)$, а потом његова Монте Карло апроксимација, назива се бутстреп метод.

Пример 10.3 (бутстреповање очекивања)**.** Претпоставимо да нас, као и у једном од претходних примера, занима оцењивање пристрасности. Овде конкретно желимо да оценимо

$$\lambda_n(F) = \mathbf{E}_F \delta(X_1, \dots, X_n) - h(F)$$

за неку оцену δ неког функционала $\theta = h(F)$. Тада је

$$\lambda_n(\hat{F}_n) = \mathbf{E}_{\hat{F}_n} \delta(X_1^*, \dots, X_n^*) - h(\hat{F}_n).$$

Видели смо да проблем настаје у рачунању умањеника у горњем изразу, јер је то рачунање за велике n рачунски веома захтевно. Да бисмо превазишли овај проблем, вучемо бутстреп узорак као и раније. За свако i одређујемо

$$\delta_i^* = \delta(X_{i1}^*, \dots, X_{in}^*),$$

а онда апроксимирамо $\lambda_n(\hat{F}_n)$ помоћу

$$\lambda_{B,n}^* = \frac{1}{B}\sum_{i=1}^{B} \delta_i^* - h(\hat{F}_n).$$

На основу закона великих бројева, умањеник у горњем изразу, као сума IID случајних величина, тежи у вероватноћи ка

$$\mathbf{E}\delta_i^* = \mathbf{E}\delta(X_1^*, \dots, X_n^*),$$

те је стога, за велико B, $\lambda_{B,n}^*$ са великом вероватноћом блиско вредности $\lambda_n(\hat{F}_n)$.
Сличан приступ може се користити и када се оцењује неки други параметар.

Постоји још много ствари које се са бутстрепом могу радити: интервали поверења (тачан, Валдов, перцентилни, стожерни, bias-corrected, Студентизован итд), затим тестирање хипотеза, поређење са другим методима, али сматрам да би то превише премашило оквире овог курса, те ћемо се овде зау-

ставити. За оне који желе да знају више, препоручује се [7], [4], [6] и слична литература.

„Боље је знати много и често, него не знати ништа, повремено, никад, којекуде.”

- Тетак Ђорђије (Црни Груја)

Библиографија

[1] Ленка Главаш (2022). *Предавања из Увода у теорију узорака*. Математички факултет, Београд.

[2] Марко Обрадовић, Марија Цупарић (2022). *Предавања и вежбе из Одабраних поглавља математичке статистике*. Математички факултет, Београд.

[3] Милан Јовановић (2020). *Предавања из Теорије узорака*. Математички факултет, Београд.

[4] D.D. Boos, L.A. Stefanski (2013). *Essential Statistical Inference: Theory and Methods*. Springer Texts in Statistics, Springer.

[5] Мато Пижурица, Митар Пешикан, Јован Јерковић (2010). *Правопис српскога језика*. Матица српска, Нови Сад.

[6] L. Wasserman (2006). *All of Nonparametric Statistics*. Springer Science & Business Media.

[7] E. Lehmann (2004). *Elements of Large-Sample Theory*. Springer Science & Business Media.