Univerzitet u Beogradu

Matematički fakultet

Seminarski rad iz predmeta Teorija uzoraka

Tema:

Ilustracija principa uzorkovanja u istraživanju o prosečnoj starosti epileptičara

Profesor:

Jelena Jocković

Asistent: Student:

Bojana Todić Jovana Protić 162/2012

Beograd, april 2016. godine

Sadržaj:

1. UVOD

Thall i Vail (1990) su dali dvonedeljni proračun za 59 epileptičara. Broj napada je bio zabeležen za osnovni period od 8 nedelja, i pacijenti su na slučajan način bili podeljeni u dve grupe: grupu za lečenje i kontrolnu grupu. Proračun je beležen 4 puta po 2 nedelje. Jedina kovarijanta su bile njihove godine.

Naša baza sadrži 236 vrta i 9 kolona. Promenljive koje se u njoj nalaze su:

1. **Y –** proračun za period od 2 nedelje
2. **trt –** tretman: placebo ili progabide
3. **basa –** proračun za osnovni period od 8 nedelja
4. **age** – starost, računata u godinama
5. **V4** – 0/1 indikator perioda od 1 do 4
6. **subject** – od 1 do 59
7. **period** – period od 1 do 4
8. **lbase** – logaritamski proračun za osnovni period
9. **lage** – logaritamski proračun starosti

Prvo ćemo se malo bolje upoznati sa našom bazom, a zatim ćemo odrediti cilj istraživanja.

```
> data(epil)
> summary(epil)
       y                  trt           base
 Min.   :  0.000   placebo  :112   Min.   :  6.00
 1st Qu.:  2.750   progabide:124   1st Qu.: 12.00
 Median :  4.000                   Median : 22.00
 Mean   :  8.254                   Mean   : 31.22
 3rd Qu.:  9.000                   3rd Qu.: 41.00
 Max.   :102.000                   Max.   :151.00
      age              V4          subject        period
 Min.   :18.00   Min.   :0.00   Min.   : 1   Min.   :1.00
 1st Qu.:23.00   1st Qu.:0.00   1st Qu.:15   1st Qu.:1.75
 Median :28.00   Median :0.00   Median :30   Median :2.50
 Mean   :28.34   Mean   :0.25   Mean   :30   Mean   :2.50
 3rd Qu.:32.00   3rd Qu.:0.25   3rd Qu.:45   3rd Qu.:3.25
 Max.   :42.00   Max.   :1.00   Max.   :59   Max.   :4.00
     lbase               lage
 Min.   :-1.36249   Min.   :-0.42941
 1st Qu.:-0.66934   1st Qu.:-0.18429
 Median :-0.06321   Median : 0.01242
 Mean   : 0.00000   Mean   : 0.00000
 3rd Qu.: 0.55932   3rd Qu.: 0.14595
 Max.   : 1.86303   Max.   : 0.41789
>
```

Iz priloženog vidimo da najmlađi ispitanik ima 18 godina, dok najstariji ima 42. Takođe vidimo, da, na placebo tretmanu ima 112 osoba, dok na progabide ima 124. Ovo možemo i grafički predtsaviti:

Sada ćemo analizirati prosečnu starost, tj izračunaćemo osnovne statistike:

```
> mean(epil$age)
[1] 28.33898
> var(epil$age)
[1] 39.20375
```

Možemo analizirati i starost u odnosu na neke kategorije, na primer, u odnosu na period posmatranja ili tip tretmana.

Kada posmatramo u odnosu na period:

```
> mean(epil.period$`1`$age)
[1] 28.33898
> mean(epil.period$`2`$age)
[1] 28.33898
> mean(epil.period$`3`$age)
[1] 28.33898
> mean(epil.period$`4`$age)
[1] 28.33898
```

Vidimo da dobijamo iste vrednosti, što je i logično, jer broj ispitanika ostaje isti.

Kada posmatramo u odnosu na indikator V4:

```
> mean(epil.v4$`0`$age)
[1] 28.33898
> mean(epil.v4$`1`$age)
[1] 28.33898
```

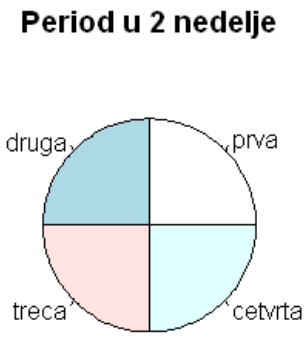
Takođe dobijamo iste vrednosti.

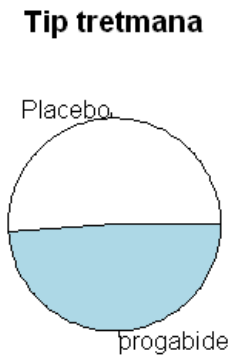
A kada posmatramo u odnosu na tretman:

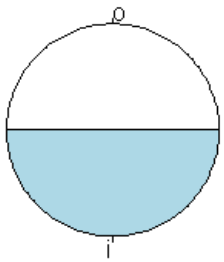
```
> mean(epil.trt$placebo$age)
[1] 29
> mean(epil.trt$progabide$age)
[1] 27.74194
```

Ovi rezultati se razlikuju, jer nije isti broj ispitanika u oba tipa tretmana.

Grafički prikaz dobijenih rezultata:







Dakle, na placebo tretmanu ima 112 osoba sa prosečnom starošću od 29 godina, dok na progabide tretmanu ima 124 osobe sa prosečnom starošću od 27,74194 godina.

Pored ovih ¨pie¸¸ grafika, zavisnost možemo prikazati i boxpolotovima, plotovima...

Kod u R-u:

```
install.packages("MASS")
library(MASS)
data(epil)
epil
summary(epil)
attach(epil)
epil$age
pie(c(112,124), labels=c("placebo", "progabide"), col=c("white","blue"))
mean(epil$age)
var(epil$age)

#odnos prema periodu
epil.period<-split(epil, epil$period,drop=FALSE)
#odnos prema vrsti lecenja
epil.trt<-split(epil,epil$trt,drop=FALSE)
mean(epil.period$`1`$age)
mean(epil.period$`2`$age)
mean(epil.period$`3`$age)
mean(epil.period$`4`$age)
mean(epil.trt$placebo$age)
mean(epil.trt$progabide$age)
```

pie(c(mean(epil.period$`1`$age),mean(epil.period$`2`$age),mean(epil.period$`3`$age),mean(epil.period$`4`$age),labels=c("prva","druga","treca","cetvrta", main="Period po 2 nedelje"))

pie(c(mean(epil.trt$placebo$age),mean(epil.trt$progabide$age)),labels=c("Placebo","progabide"),main="Tip tretmana")

2. CILJ ISTRAŽIVANJA

Sada kada smo se malo bolje upoznali sa našom bazom, sada možemo i postaviti cilj istraživanja. Dakle, odgovorićemo na pitanja o prosečnoj starosti na placebo tretmanu, kao i na progabide tretmanu.

3. OBIM UZORKA

Obim ćemo računati po formuli: $\eta = \frac{1}{\frac{1}{\eta_0}+\frac{1}{N}}$, $\eta_0 = \frac{z^2 N^2 S_y^2}{d^2}$, gde vrednost za z čitamo iz tablica za normalnu raspodelu i ona iznosi z=1,96 za 95% interval poverenja.

Obim populacije je N=236. Disperziju S_y^2 ocenjujemo iz uzorka manjeg obima, kao i srednju vrednost koju ćemo koristiti za procenu greške. Uzećemo da je d=0,1*mean(age).

Kod u R-u:

N=nrow(epil)
N
z=1.96
s<-sort(sample(length(epil[,1]),150))
uzorak<-epil[s,]
m<-mean(epil$age)
v<-var(uzorak$age)
*d<-0.1*m*
*n0<-(z^2*v*N^2)/d^2*
n<-1/(1/n0+1/N)
n

Međutim, ovo nam daje vrednost od 235.9453 što i nije dobar rezultat.

Sad sam disperziju smanjila za 10% i u daljem radu sam računala da je obim jednak 20.

m<-mean(epil$age)
d<-0.1*m

```
uzorak<-sample(epil$age,24)
S<-sqrt(var(uzorak))
n0<-1/(1/N+d^2/(z^2*S^2))
n<-1/(1/n0+1/N)
n
```

4.PROST SLUČAJAN UZORAK

Prost slučajan uzorak bez ponavljanja je plan po kome se n različitih jedinica uzorka bira na takav način da svaka moguća kombinacija od n jedinica ima istu verovatnoću da bude izabrana iz populacije.

Kod u R-u:

```
PSU=epil[sample(1:nrow(epil),n,replace=FALSE),]
ypsu=sum(PSU$age)/n #ocena sredine obelezja
ypsu
Ypsu=N*sum(PSU$age)/n #Horvitz-Thompson-ova ocena totala
Ypsu
#sada racunamo korigovanu uzoracku disperziju
spsu=0
for(i in 1:n) spsu=spsu+(PSU$age[i]+ypsu)
spsu=spsu/(n-1)
spsu
Vypsu=(N-n)*spsu/(N*n) #ocena disperzije sredine obelezja
VYpsu=N*(N-n)*spsu/n #ocena disperzije totala
VYpsu
```

Dobijeni su sledeći rezultati: ocena sredine obeležja(ypsu)=29,75; ocena totala(Ypsu)=7021; disperzija sredine obeležja(Vypsu)=2,866; disperzija totala(VYpsu)=159635,4.

5.STRATIFIKOVANI SLUČAJNI UZORAK

Stratifikovani uzorak primenjuje se kada se treba povećati preciznost ocena parametara, tj smanjiti greške uzorka. Stratifikacija je podela populacije na potpopulacije, stratume.

Neke od oznaka koje ćemo koristiti:

- L – broj stratuma u populaciji

- N_h - broj jedinica u h-tom stratumu, h=1,...,L
- n_h - veličina uzorka koji se bira iz h-tog stratuma

Postoje dve tehnike za odrađivanje obima uzorka n_h za svaki stratum pojedinačno:

- ❖ Proporcionalni raspored
- ❖ Optimalni raspored

5.1. PROPORCIONALNI RASPORED

Kod proporcionalnog rasporeda, broj jedinica koje se biraju u uzorak iz pojedinog stratuma, proporcionalan je broju jedinica u tom stratumu, tj. $n_h = \frac{n}{N} N_h,\ h = 1, \dots, L.$

Naš uzorak delimo na stratume prema vrsti tretmana. U našem slučaju h uzima vrednosti 1 ili 2, jer imamo samo dve podele: placebo ili progabide. Sada računamo obime stratuma prema već navedenoj formuli:

*n1=n/N*nrow(epil.trt$placebo)*
*n2=n/N*nrow(epil.trt$progabide)*

Sada, kada smo odredili obime, možemo da izdvojimo odgovarajuće stratume:

L1=epil.trt$placebo[sample(1:nrow(epil.trt$placebo),n1,replace=FALSE),]
L2=epil.trt$progabide[sample(1:nrow(epil.trt$progabide),n2,replace=FALSE),]

```
> summary(L1)
       y                trt          base             age              V4
 Min.   : 0.00   placebo  :69   Min.   :  6.00   Min.   :19.00   Min.   :0.0000
 1st Qu.: 3.00   progabide: 0   1st Qu.: 11.00   1st Qu.:24.00   1st Qu.:0.0000
 Median : 4.00                  Median : 20.00   Median :29.00   Median :0.0000
 Mean   : 8.87                  Mean   : 31.22   Mean   :28.83   Mean   :0.2609
 3rd Qu.:12.00                  3rd Qu.: 47.00   3rd Qu.:32.00   3rd Qu.:1.0000
 Max.   :76.00                  Max.   :111.00   Max.   :42.00   Max.   :1.0000
    subject          period         lbase               lage
 Min.   : 1.00   Min.   :1.000   Min.   :-1.36249   Min.   :-0.37534
 1st Qu.: 9.00   1st Qu.:2.000   1st Qu.:-0.75635   1st Qu.:-0.14173
 Median :16.00   Median :2.000   Median :-0.15852   Median : 0.04751
 Mean   :15.58   Mean   :2.522   Mean   :-0.02456   Mean   : 0.01937
 3rd Qu.:22.00   3rd Qu.:4.000   3rd Qu.: 0.69590   3rd Qu.: 0.14595
 Max.   :28.00   Max.   :4.000   Max.   : 1.55528   Max.   : 0.41789
> summary(L2)
       y                 trt          base             age              V4
 Min.   :  0.000   placebo  : 0   Min.   :  7.00   Min.   :18.00   Min.   :0.0000
 1st Qu.:  2.000   progabide:76   1st Qu.: 14.00   1st Qu.:22.00   1st Qu.:0.0000
 Median :  4.000                  Median : 24.00   Median :26.00   Median :0.0000
 Mean   :  7.579                  Mean   : 32.49   Mean   :27.79   Mean   :0.2763
 3rd Qu.:  8.000                  3rd Qu.: 41.00   3rd Qu.:33.50   3rd Qu.:1.0000
 Max.   :102.000                  Max.   :151.00   Max.   :41.00   Max.   :1.0000
    subject          period         lbase               lage
 Min.   :29.00   Min.   :1.000   Min.   :-1.20834   Min.   :-0.42941
 1st Qu.:38.75   1st Qu.:1.000   1st Qu.:-0.51519   1st Qu.:-0.22874
 Median :44.00   Median :3.000   Median : 0.02380   Median :-0.06169
 Mean   :44.46   Mean   :2.513   Mean   : 0.08056   Mean   :-0.02467
 3rd Qu.:50.25   3rd Qu.:4.000   3rd Qu.: 0.55932   3rd Qu.: 0.19143
 Max.   :59.00   Max.   :4.000   Max.   : 1.86303   Max.   : 0.39379
```

Sada imamo sve što nam je potrebno da bismo odredili ocenu sredine populacije, ocenu totala, kao i njihove varijanse.

Kod u R-u:

```
#sada racunamo ocenu sredine obelezja populacije:
#prvo racunamo srednje vrednosti prvog i drugog stratuma
y1=sum(L1$age)/n1
y2=sum(L2$age)/n2
#ocena sredine obelezja:
N1=112
N2=124
yst=(N1*y1+N2*y2)/N
yst
#sada racunamo varijansu ocene sredine
s1=0
s2=0
for(i in 1:n1) s1=s1+((L1$age[i]-y1)^2)
s1=s1/(n1-1)
for(j in 1:n2) s2=s2+((L2$age[j]-y2)^2)
s2=s2/(n2-1)
Vyst=(N1*(N1-n1)*s1/n1 + N2*(N2-n2)*s2/n2)/(N^2)
Vyst
#sada racunamo ocenu totala populacije, kao i njene varijanse
Yst=N1*y1+N2*y2
Yst
#ocena njegove varijanse
VYst=N1*(N1-n1)*s1/n1 + N2*(N2-n2)*s2/n2
VYst
```

Dobijeni su sledeći rezultati: ocena sredine obeležja(yst)=28,8; ocena totala(Yst)=6088,8; disperzija sredine obeležja(Vyst)=1,0708; disperzija totala(VYst)=59643,58.

5.2. OPTIMALNI RASPORED

Prethodno opisana tehnika proporcionalnog rasporeda ne uzima u obzir nijedan drugi aspekt predmeta istraživanja osim veličine stratuma. Ona u potpunosti ignoriše unutrašnju strukturu stratuma. Zato su predložene šeme rasporeda. U praksi se koriste dve šeme rasporeda koje minimiziraju disperziju ocena. Kako je minimalna disperzija optimalno svojstvo ocene, ovakvi rasporedi se nazivaju optimalni. Šeme su:

- Neyman-ov raspored
- Cost Optimum Allocation

Sada ćemo naše stratume izdvajati u odnosu na indikator V4, pri čemu ćemo koristiti Neyman-ovu šemu. I u ovoj podeli h uzima vrednosti 1 ili 2, jer indikator V4 može da uzme vrednosti 0 ili 1. Računamo broj jedinica po stratumima:

```
> N1=nrow(epil.v4$`0`)
> N2=nrow(epil.v4$`1`)
> N1
[1] 177
> N2
[1] 59
```

Vidimo da je broj jedinica za V4=0 jednak 177, a kada je V4=1 iznosi 59.

```
#racunamo odgovarajuce varijanse
s1=var(epil.v4$`0`$age)
s2=var(epil.v4$`1`$age)
s1
s2
#odredjujemo n1 i n2 koji minimizira disperziju
n1=n*s1*N1/(N1*s1+N2*s2)
n2=n*N2*s2/(N1*s1+N2*s2)
n1
n2
#izdvajamo odgovarajuce stratume
L1=epil.v4$`0`[sample(1:nrow(epil.v4$`0`),n1,replace=FALSE),]
L2=epil.v4$`1`[sample(1:nrow(epil.v4$`1`),n2,replace=FALSE),]
```

Sada možemo da izdvojimo odgovarajuće stratume:

```
> summary(L1)
       y                 trt          base             age              V4
 Min.   :  0.000   placebo  :56   Min.   :  7.00   Min.   :18.00   Min.   :0
 1st Qu.:  2.000   progabide:53   1st Qu.: 12.00   1st Qu.:24.00   1st Qu.:0
 Median :  4.000                  Median : 22.00   Median :28.00   Median :0
 Mean   :  8.202                  Mean   : 29.01   Mean   :28.92   Mean   :0
 3rd Qu.:  8.000                  3rd Qu.: 38.00   3rd Qu.:33.00   3rd Qu.:0
 Max.   :102.000                  Max.   :151.00   Max.   :42.00   Max.   :0
    subject          period          lbase              lage
 Min.   : 1.00   Min.   :1.000   Min.   :-1.20834   Min.   :-0.42941
 1st Qu.:13.00   1st Qu.:1.000   1st Qu.:-0.66934   1st Qu.:-0.14173
 Median :27.00   Median :2.000   Median :-0.06321   Median : 0.01242
 Mean   :29.21   Mean   :1.963   Mean   :-0.05438   Mean   : 0.02020
 3rd Qu.:44.00   3rd Qu.:3.000   3rd Qu.: 0.48334   3rd Qu.: 0.17672
 Max.   :59.00   Max.   :3.000   Max.   : 1.86303   Max.   : 0.41789
> summary(L2)
       y                 trt          base             age              V4
 Min.   : 0.000   placebo  :21   Min.   :  8.00   Min.   :18.00   Min.   :1
 1st Qu.: 3.000   progabide:15   1st Qu.: 15.25   1st Qu.:24.00   1st Qu.:1
 Median : 5.000                  Median : 25.50   Median :28.50   Median :1
 Mean   : 7.222                  Mean   : 32.33   Mean   :28.72   Mean   :1
 3rd Qu.: 8.250                  3rd Qu.: 43.00   3rd Qu.:33.50   3rd Qu.:1
 Max.   :29.000                  Max.   :111.00   Max.   :42.00   Max.   :1
    subject          period          lbase              lage
 Min.   : 4.00   Min.   :4   Min.   :-1.07481   Min.   :-0.42941
 1st Qu.:13.50   1st Qu.:4   1st Qu.:-0.43357   1st Qu.:-0.14173
 Median :24.50   Median :4   Median : 0.08270   Median : 0.02997
 Mean   :26.86   Mean   :4   Mean   : 0.09042   Mean   : 0.01370
 3rd Qu.:39.25   3rd Qu.:4   3rd Qu.: 0.60616   3rd Qu.: 0.19143
 Max.   :58.00   Max.   :4   Max.   : 1.55528   Max.   : 0.41789
```

Kada smo sve ovo odredili, sada možemo izračunati ocenu srednje vrednosti obeležja, totala, kao i njihovih varijansi.

Kod u R-u:

```
#sada racunamo ocenu srednje vrednosti obelezja i njenu varijansu
y1=sum(L1$age)/n1 #srednja vrednost prvog stratuma
y2=sum(L2$age)/n2 #srednja vrednost drugog stratuma
yop=(N1*y1+N2*y2)/N
yop
#sada ocenjujemo varijansu
s1=0
s2=0
for(i in 1:n1) s1=s1+((L1$age[i]-y1)^2)
s1=s1/(n1-1)
for(j in 1:n2) s2=s2+((L2$age[jj]-y2)^2)
s2=s2/(n2-1)
Vyop=(N1*(N1-n1)*s1/n1 + N2*(N2-n2)*s2/n2)/(N^2)
Vyop
#jos je ostalo da ocenimo total
Yop=N1*y1+N2*y2
Yop
#varijansa
VYop=N1*(N1-n1)*s1/n1 + N2*(N2-n2)*s2/n2
VYop
```

Dobijeni su sledeći rezultati: ocena sredine obeležja(yop)=27,94; ocena totala(Yop)=6594,58; disperzija sredine obeležja(Vyop)=2,1626; disperzija totala(VYop)=120452,5.

6. SISTEMATSKI UZORAK

Kod sistematskog uzorka, umesto izbora n jedinica iz populacije na slučajan način, odlučivanje o jedinicama koje će se naći u uzorku vrši na osnovu izbora samo jednog slučajnog broja.

Prvo ćemo napraviti uzorak:

```
#sistematski uzorak
k=N/n
k
k=2
r<-sample(1:k, 1)
#sada formiramo uzorak
v<-seq(r, r+(n-1)*k,k)
v
P<-epil$age[v]
P
```

Sada možemo izračunati ocene sredine obeležja i totala, kao i njihove disperzije.

Kodu R-u:

```
k=round(N/n)
k
r<-sample(1:k, 1)
#sada formiramo uzorak
v<-seq(r, r+(n-1)*k,k)
v
P<-epil$age[v]
P
#kada smo formirali uzorak, racunamo ocenu sredine i totala
#ocena sredine
ys=sum(P)/n
ys
#ocena totala
Ys=N*sum(P)/n
#disperzije
ss=0
for(i in 1:n) ss=ss+(P[i]+ys)^2
ss=ss/(n-1)
Vys=(N-n)*ss/(N*n)
Vys
VYs=N*(N-n)*ss/n
VYs
```

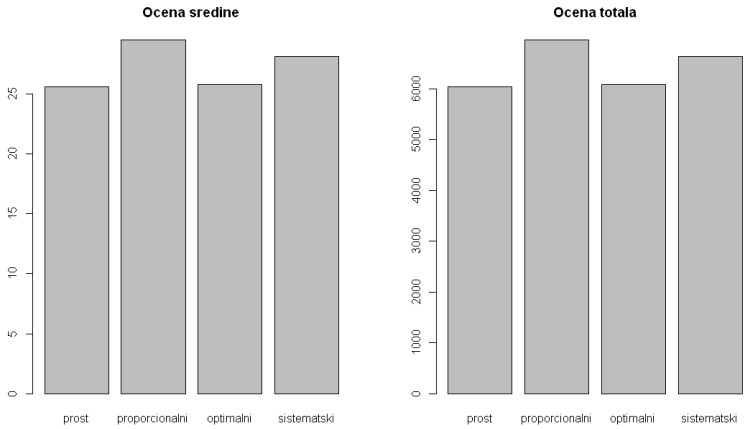
Dobijeni su sledeći rezultati: ocena sredina(ys)=28,1; ocena totala(Ys)=6631,6; disperzija sredine(Vys)=153,34; disperzija totala(VYs)=8540439.

6.ZAKLJUČAK

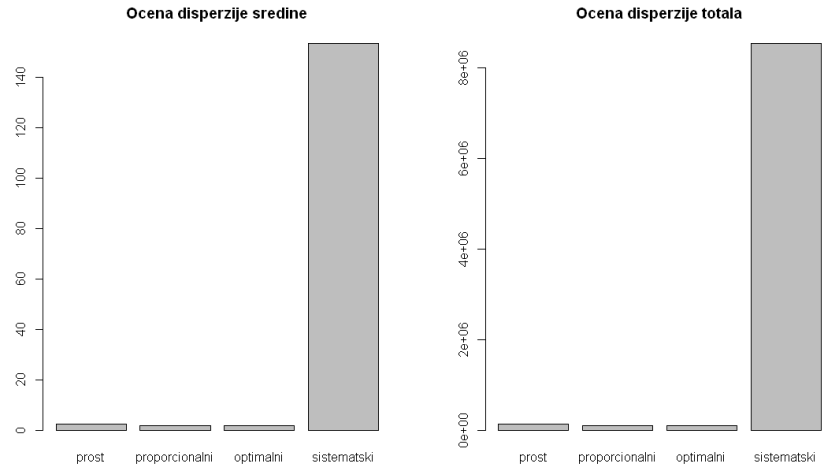
Uzorak:	Ocena sredine obeležja:	Disperzija sredine obeležja:	Ocena totala:	Disperzija totala:
Prost	29,75	7021	2,866	159635,4
Proporcionalni	28,8	6088,8	1,0708	59643,58
Optimalni	27,94	6594,58	2,1626	120452,5
Sistematski	28,1	6631,6	153,34	8560639

Sada ćemo grafički uporediti planove uzorkovanja na osnovu određenog obima:

Prvo ćemo uporediti dobijene ocene sredine i totala:



Zatim ćemo uporediti njihove disperzije:



Kod u R-u:

```
par(mfrow=c(1,2))

barplot(c(ypsu, yst, yop, ys),names.arg = c("prost", "proporcionalni", "optimalni", "sistematski"),
main="Ocena sredine")
barplot(c(Ypsu, Yst, Yop, Ys),names.arg = c("prost", "proporcionalni", "optimalni",
"sistematski"), main="Ocena totala")
#uporedjujemo dobijene disperzije
barplot(c(Vypsu, Vyst, Vyop, Vys),names.arg = c("prost", "proporcionalni", "optimalni",
"sistematski"), main="Ocena disperzije sredine")
barplot(c(VYpsu, VYst, VYop, VYs),names.arg = c("prost", "proporcionalni", "optimalni",
"sistematski"), main="Ocena disperzije totala")
```

Na osnovu dobijenih rezultata disperzije možemo zaključiti da i prost, kao i stratifikovani uzorak (i proporcionalni i optimalni) daju bolje rezultate od sistematskog uzorka, za obim jednak 20.

Literatura:

- http://www.matf.bg.ac.rs/p/bojana-todic
- https://www.facebook.com/l.php?u=https%3A%2F%2Fedis.ifas.ufl.edu%2Fpd006&h=cAQG1_6ZM&s=1