Analiza uspeha učenika srednjih škola u SAD

Seminarski rad u okviru kursa
Uvod u teoriju uzoraka
Matematički fakultet

Aleksandar Jakovljević
mi15156@alas.matf.bg.ac.rs

1. juli 2019

Sažetak

U radu je izvršena analiza uspeha učenika srednjih škola u SAD. Izvršeno je poređenje između planova izbora uzoraka i analizirani su njihovi rezultati. Rezultati su pokazali da stratifikovan slučajan uzorak daje bolju ocenu srednje vrednosti populacije u odnosu na prost slučajan uzorak bez ponavljanja i predstavlja bolji plan izbora uzorka nad zadatom populacijom.

Sadržaj

1 Uvod

Rezultati učenika srednjih škola su značajni radi procenjivanja socioekonomskih aspekata društva i praćenje njihovih rezulata može da ukaže na probleme ili poboljšanja u društvu i sistemu. Rezultati učenika zavise od mnogobrojnih obeležja i potrebno je analizirati na koji način su vrednosti pomenutih obeležja povezane sa rezultatima.

1.1 Opis baze podataka

Baza se sastoji od 1000 redova i 8 kolona. Redovi predstavljaju jedinice populacije, dok kolone predstavljaju obeležja. Obeležja su podeljena u dve grupe, od kojih je 5 kategoričkih i 3 numerička. Radi jednostavnijeg računa, bazi podataka smo dodali jednu kolonu koja je numerička i predstavlja prosečan rezultat sva tri testa. Od kategoričkih obeležja u bazi se javljaju sledeća:

gender obeležje koje uzima vrednosti iz skupa {male, female} i opisuje pol učenika.

race/ethnicity obeležje koje uzima vrednosti iz skupa { group A, group B, group C, group D, group E} i opisuje rasnu, odnosno etničku pripadnost.

parental level of education obeležje koje uzima vrednost iz skupa {high school, some high school, bachelor's degree, master's degree, associate's degree, some college} i opisuje stepen obrazovanja roditelja.

lunch obeležje koje uzima vrednost iz skupa {standard, free/reduced} i opisuje da li je učenik imao obroke standardne cene i veličine ili besplatne obroke redukovane veličine.

test preparation course obeležje koje uzima vrednost iz skupa {completed, none} i opisuje da li je učenik radio probni test.

Od numeričkih podataka u bazi se nalaze:

math score obeležje koje opisuje rezultat na testu iz matematike uzima vrednost iz skupa prirodnih brojeva i pripada domenu [0, 100].

reading score obeležje koje opisuje rezultat na testu iz čitanja uzima vrednost iz skupa prirodnih brojeva i pripada domenu [0, 100].

writing score obeležje koje opisuje rezultat na testu iz pisanja uzima vrednost iz skupa prirodnih brojeva i pripada domenu [0, 100].

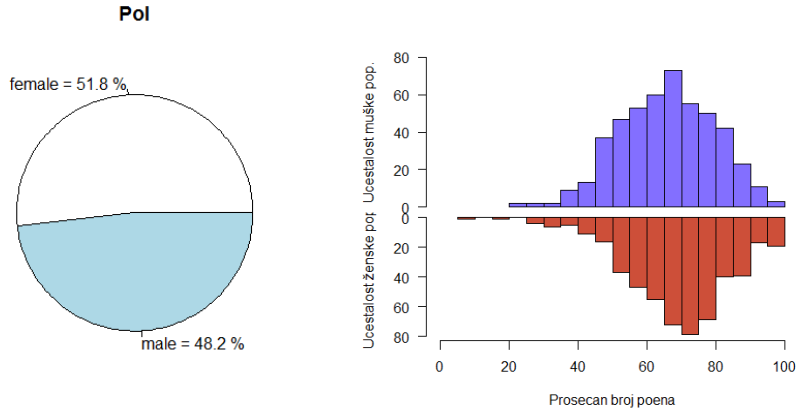
Pored ovih osam, dodato je još jedno obeležje.

total obeležje koje opisuje aritmetičku sredinu na osnovu rezultata postignutih na testu iz matematike, čitanja i pisanja koje uzima vrednost iz skupa realnih brojeva i pripada domenu [0, 100].

Bazu podataka možete pronaći na Kaggle: Students Performance in Exams.

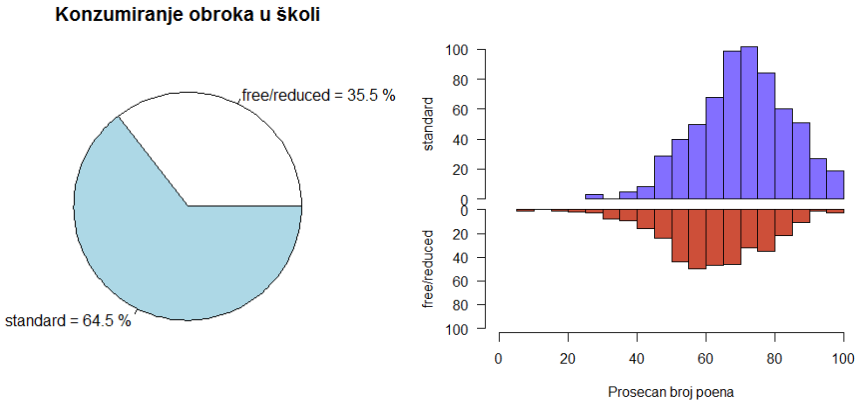
2 Analiza baze podataka

Pre nego što počnemo rad sa bazom, neophodno je izvršiti analizu podataka i imati predstavu sa kakvim podacima radimo. Na slici 1 možemo uočiti udeo polova (obeležje *gender*) unutar populacije i ostvarene rezultate prema polovima. Većinu populacije čine osobe ženskog pola 1a i može se uočiti sa histograma da su osobe ženskog pola ostvarile bolji



(a) Udeo polova u populaciji (b) Histogrami učestalosti obeležja *total*

Slika 1: Analiza obeležja *gender*.

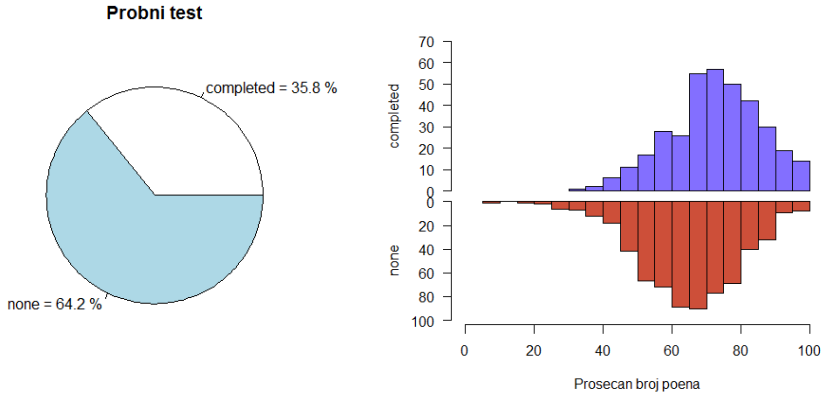


(a) Udeo konzumiranja obroka u populaciji (b) Histogrami učestalosti obeležja *total*

Slika 2: Analiza obeležja *lunch*.

prosečan rezultat u odnosu na osobe muškog pola. Ukoliko dalje posmatramo obeležje *lunch* 2, možemo uočiti da približno dve trećine populacije čine učenici koji imaju standardne obroke 2a. Na histogramu se jasno vidi da su rezultati učenika sa boljom ishranom bolji od učenika koji imaju redukovanu ishranu 2b. Analizom obeležja *test preparation*, čiji je udeo dat na slici 3a, možemo sa histograma 3b videti da su studenti koji su odradili probni test postigli bolje rezultate.

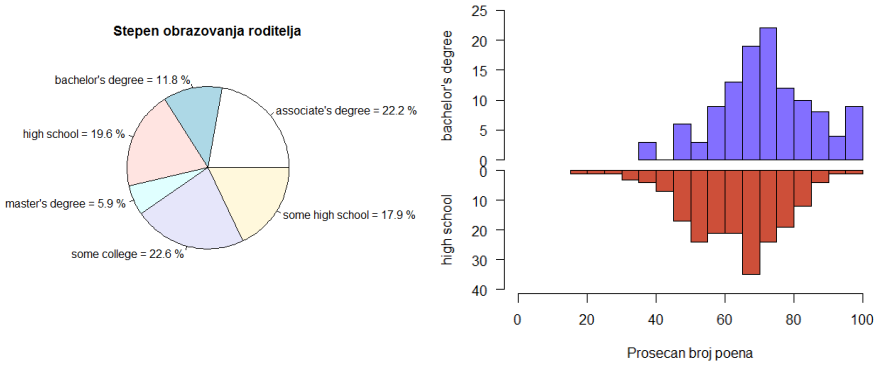
Kardinalnost skupova prethodnih obeležja je iznosila 2, dok kardinalost skupova obeležja *parental level of education* i *race/ethnicity* iznosi 5 i 6, respektivno. Na slikama 4a i 5a su dati udeli vrednosti obeležja *parental level of education* i *race/ethnicity* unutar populacije.
Kako bismo lakše predstavili učestalost obeležja u odnosu na postignute rezultate, iz skupa oba obeležja, uzećemo po dve vrednosti i uporediti ih. Na histogramu 4b možemo uočiti da su rezultati učenika čiji su roditelji



(a) Udeo odrađenog probnog testa u populaciji

(b) Histogrami učestalosti obeležja *total*

Slika 3: Analiza obeležja *test preparation*.



(a) Udeo stepena obrazovanja u populaciji

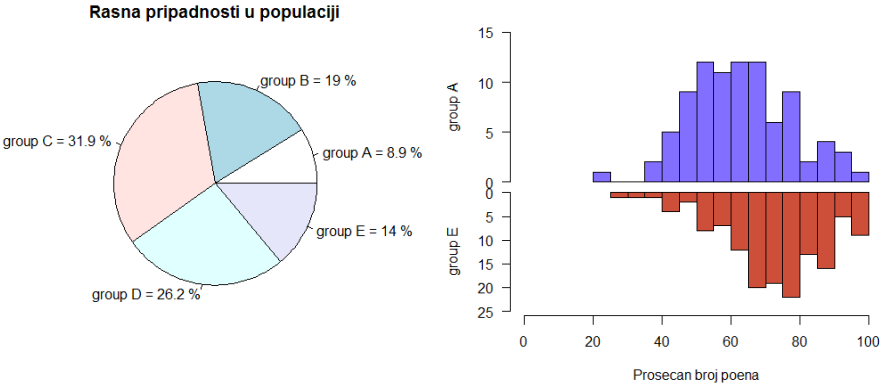
(b) Histogrami učestalosti obeležja *total*

Slika 4: Analiza obeležja *parental level of education*.

završili fakultet (engl. *bachelor's degree*) učestaliji na intervalu [90, 100], kao i da je prosek veći. U slučaju obeležja *race/ethnicity*, sa histograma 5b jasno se vidi da su učenici iz grupe E postigli bolje rezultate od učenika iz grupe A. U nastavku se nalaze primeri kodova koji iscrtavaju pitu 1 i histogram 2.

```
mytable = table(podaci$lunch)
lbls = paste(names(mytable), "=",
             mytable/10,"%",
             sep=" ")
pie(mytable,
    labels = lbls,
    main="Obrok")
```

Listing 1: Primer koda kojim dobijamo dijagram pite



(a) Udeo rasne pripadnosti u populaciji (b) Histogrami učestalosti obeležja *total*

Slika 5: Analiza obeležja *race/ethnicity*.

```
1000
    par(mfrow=c(2,1))
1002 #Make the plot
    par(mar=c(0,5,3,3))
1004 hist(total_score[podaci$race.ethnicity  == "group A"],
         main="" ,
1006     xlim=c(0,100),
         ylab="group A",
1008     xlab="",
         xaxt ="n",
1010     ylim=c(0,15),
         las=1 ,
1012     col="slateblue1",
         breaks=20)
1014 par(mar=c(5,5,0,3))
    hist(total_score[podaci$race.ethnicity  == "group E"] ,
1016     main="" ,
         xlim=c(0,100),
1018     ylab="group E",
         xlab="Prosecan broj poena",
1020     ylim=c(25,0) ,
         las=1 ,
1022     col="tomato3",
         breaks=20)
```

Listing 2: Primer koda kojim dobijamo histograme

3 Teorijski uvod

Nad datom bazom podataka koristili smo dva plana izbora uzorka: prost slučajan uzorak bez ponavljanja i stratifikovan slučajan uzorak. U daljem delu rada, navešćemo neke osobine navedenih planova i formule korišćene prilikom izvršavanja planova. U poglavlju 4 su obrađeni rezultati dobijeni izvršavanjem naredna dva plana izbora uzorka.

3.1 Prost slučajan uzorak bez ponavljanja

Prost slučajan uzorak bez ponavljanja je plan izbora uzorka koji iz populacije od N jedinica, nasumično bira n različitih jedinica tako da

svaka kombinacija od n jedinica ima istu verovatnoću da bude izabrana iz populacije. Verovatnoća da određena kombinacija bude izabrana iznosi

$$p = \begin{cases} \binom{N}{n}^{-1}, & \text{ako je obim uzoraka jednak } n \\ 0, & \text{inače} \end{cases} \tag{1}$$

Neka je sa S označen prost slučajan uzorak bez ponavljanja. Tada ocena populacijske srednje vrednosti u oznaci m_Y iznosi

$$\widehat{m}_Y = \frac{1}{n} \sum_{k \in S} y_k \tag{2}$$

Ova ocena je nepristrasna, odnosno važi

$$E(\widehat{m}_Y) = m_Y \tag{3}$$

Disperziju ocene i ocenu disperzije računamo na sledeći način

$$D(\widehat{m}_Y) = \frac{\sigma^2}{n}\left(1 - \frac{n}{N}\right) \tag{4}$$

$$\widehat{D}(\widehat{m}_Y) = \frac{\bar{S}^2}{n}\left(1 - \frac{n}{N}\right) \tag{5}$$

gde je σ^2 populacijska disperzija, a $\bar{S}^2$ uzoračka disperzija. Kada su u pitanju intervalne ocene, aproksimativni $100 \cdot (1-\alpha)\%$ dvostrani interval poverenja računamo pomoću sledeće formule:

$$I_{\widehat{m}_Y} = \left[\widehat{m}_Y - z_{1-\frac{\alpha}{2}}\sqrt{\widehat{D}(\widehat{m}_Y)}, \widehat{m}_Y + z_{1-\frac{1}{\alpha}}\sqrt{\widehat{D}(\widehat{m}_Y)}\right] \tag{6}$$

gde je za $n \geq 30$, $z_{1-\frac{\alpha}{2}}$ vrednost $\left(1-\frac{\alpha}{2}\right)$-kvantila standardne normalne raspodele. Ukoliko n ne ispunjava uslov, $z_{1-\frac{\alpha}{2}}$ uzima vrednost $\left(1-\frac{\alpha}{2}\right)$-kvantila t_{n-1} raspodele i označava se sa $t_{n-1;1-\frac{\alpha}{2}}$.

Nakon što smo ukratko naveli formule koje su povezane sa prostim slučajnim uzorkom bez ponavljanja [2], u nastavku sledi kod 3 napisan u programskom jezik **R**, koji simulira ovaj plan izbora uzorka.

```
1000
n = 278 # velicina uzorka
1002 N = length(podaci$gender) # velicina baze
set.seed(42) # postavljamo seme
1004 s = sample(N, n) # vrsimo uzorkovanje n jedinica iz populacije
     velicine N
uzorak = podaci[s,] # filtriramo populaciju
1006
# ocenjene srednje vrednosti
1008 m_ocena_mat = mean(uzorak$math.score)
m_ocena_read = mean(uzorak$reading.score)
1010 m_ocena_writ = mean(uzorak$writing.score)
m_ocena_tot = mean(uzorak$total)
1012
# ocena disperzije
1014 s_2h = var(uzorak$total)
d_ocena_tot = s_2h / n * (1 - n/N)
1016
# 95% interval poverenja ocene srednje vrednosti m_ocena_tot
1018 alpha = 0.05
z = qnorm(1 - alpha/2)
1020 interval = c(m_ocena_tot - z * sqrt(d_ocena_tot),
             m_ocena_tot + z * sqrt(d_ocena_tot) )
```

Listing 3: Prost slučajan uzorak bez ponavljanja

3.2 Stratifikovan skučajan uzorak

Stratifikacija je podela populacija na potpopulacije koji se nazivaju stratumi. Stratifikacijom vršimo klasifikaciju podataka na skupove prema jednom ili više obeležja koji su nam od ranije poznati. Dobijeni skupovi su disjunktni i potrebno je da zadovoljavaju uslov pokrivenosti, odnosno važi

$$N_1 + N_2 + ... + N_L = N \tag{7}$$

gde je $N_h, \forall h = \overline{1, L}$ veličina stratuma, a N ukupna veličina populacije. Stratumi su homogeni po vrednosti obeležja koje je iskorišćeno pri podeli populacije. Uzorak veličine n dobijen iz stratifkovane populacije sadrži iz svakog stratuma uzorak veličine $n_h, \forall h = \overline{1, L}$, gde važi da je

$$n_1 + n_2 + ... + n_L = n \tag{8}$$

Izbor jedinica iz stratuma je prost slučajan bez ponavljanja. Da bismo mogli da izračunamo ocenu srednje vrednosti obeležja *total* čitave populacije, potrebno je da ocenimo populacijski total obeležja *total* i dobijamo ga primenom naredne formule:

$$\hat{t}_Y^{str} = \sum_{h=1}^{L} N_h \cdot \widehat{m}_h = \sum_{h=1}^{L} \sum_{k \in S} \frac{n_h}{N_h} \cdot y_{hk} \tag{9}$$

gde je S uzorak, y_{hk} vrednost obeležja jedinice iz uzorka. Nakon što smo ocenili populacijski total, ocenu srednje vrednosti dobijamo iz sledeće formule:

$$\widehat{m}_y^{str} = \frac{\hat{t}_Y^{str}}{N} \tag{10}$$

Ocena je nepristrasna, odnosno važi da je

$$E(\widehat{m}_y^{str}) = m_y \tag{11}$$

Disperzija ocene i ocena disperzije se dobijaju, respektivno:

$$D(\widehat{m}_Y^{str}) = \sum_{h=1}^{L} \Big(\frac{N_h}{N}\Big)^2 \cdot \frac{\sigma_h^2}{n_h}\Big(1 - \frac{n_h}{N_h}\Big) \tag{12}$$

$$\widehat{D}(\widehat{m}_Y^{str}) = \sum_{h=1}^{L} \Big(\frac{N_h}{N}\Big)^2 \cdot \frac{S_h^2}{n_h}\Big(1 - \frac{n_h}{N_h}\Big) \tag{13}$$

gde je σ^2 populacijska disperzija, a $\bar{S}^2$ uzoračka disperzija. Na sličan način, kao i kod prostog slučajnog uzorka, računamo intervalnu ocenu, odnosno aproksimativni $100 \cdot (1 - \alpha)\%$ dvostrani interval poverenja:

$$I_{\widehat{m}_Y^{str}} = \Big[\widehat{m}_Y^{str} - z_{1-\frac{\alpha}{2}} \sqrt{\widehat{D}(\widehat{m}_Y^{str})}, \widehat{m}_Y^{str} + z_{1-\frac{1}{\alpha}} \sqrt{\widehat{D}(\widehat{m}_Y^{str})}\Big] \tag{14}$$

gde je za $n \geq 30$, $z_{1-\frac{\alpha}{2}}$ vrednost $\Big(1 - \frac{\alpha}{2}\Big)$-kvantila standardne normalne raspodele. Ukoliko n ne ispunjava uslov, $z_{1-\frac{\alpha}{2}}$ uzima vrednost $\Big(1 - \frac{\alpha}{2}\Big)$-kvantila t_{n-1} raspodele i označava se sa $t_{n-1;1-\frac{\alpha}{2}}$.

Nakon što smo ustanovili koje formule je potrebno koristiti [2], ostaje još da utvrdimo na koji način biramo broj jedinica koje si biraju u uzorak iz svakog stratuma, odnosno n_h. U ovom istraživanju korišćen je proporcionalni raspored gde važi

$$\frac{n_h}{n} = \frac{N_h}{N} \tag{15}$$

Odatle sledi da je

$$n_h = n \cdot \frac{N_h}{N} \tag{16}$$

Za verovatnoću uključenja prvog reda važi da je

$$\pi_{hk} = \frac{n}{N}, \forall h = \overline{1, L} \wedge \forall k = \overline{1, N_h} \tag{17}$$

U nastavku sledi primer koda koji prikazuje na koji način je implementiran stratifikovan slučajan uzorak u programskom jeziku **R**. U primeru koda 4, kao obeležje za podelu na stratume je uzeto obeležje *gender*. Celokupan kod možete pronaći na sledećoj veb adresi GitHub: AlexJakovljevic.

```
N_male = length(podaci[podaci$gender == "male",]$gender)
N_female = N - N_male
N_h_pol = c(N_male, N_female)
n_h_pol = N_h_pol * n / N
set.seed(42)
male = podaci[podaci$gender == "male", ][sample(1:N_male, n_h_pol
    [1]),]
male
set.seed(42)
female = podaci[podaci$gender == "female", ][sample(1:N_female, n_
    h_pol[2]),]
female

m_h_tot_pol = c(mean(male$total), mean(female$total))
t_h_tot_pol = sum(m_h_tot_pol * c(N_male, N_female))
m_h_tot_str_pol = t_h_tot_pol / N

s_2h_str_pol = c(var(male$total), var(female$total))
d_ocena_tot_str_pol = sum((N_h_pol/N)^2 * s_2h_str_pol/n_h_pol * (
    1 - n_h_pol/N_h_pol))

interval_str = c(m_h_tot_str_pol - z* sqrt(d_ocena_tot_str_pol),
                 m_h_tot_str_pol + z* sqrt(d_ocena_tot_str_pol))
```

Listing 4: Stratifikovan slučajan uzorak

4 Rezultati uzorkovanja

Za izbor veličine uzorka, konsultovana je [1]. Veličina uzorka je $n = 278$. Veličina populacije je $N = 1000$. U narednim tabelama, nalaze se rezultati dobijeni primenom odgovarajućeg plana, gde skraćenica PSU predstavlja prost slučajan uzorak bez ponavljanja, a S(obeležje) predstavlja stratifikovan slučjan uzorak na osnovu kojeg obeležja su napravljeni stratumi. U tabeli 1 možemo uočiti da je najmanju ocenu disperzije srednje vrednosti imao stratifikovan slučajan uzorak po obeležju *race/ethnicity* (*race* u tabeli). Srednja vrednost obeležja *math.score* populacije iznosi $m_Y = 66.089$.

U tabeli 2 se nalaze rezultati dobijeni nad obeležjem *reading.score*. Stratifikovan slučajan uzorak je za svako obeležje imao manju ocenu disperzije srednje vrednosti u odnosu na PSU. Najmanju ocenu disperzije srednje vrednosti je imao stratifikovan slučajan uzorak nad obeležjem *gender*. Srednja vrednost obeležja *reading.score* populacije iznosi $m_Y = 69.169$.

Naredna tabela 3 nam govori o rezultatima nad obeležjem *writing.score*. Kao i u prethodnoj tabeli, stratifikovan slučajan uzorak je pokazao bolje rezultate od PSU. Najmanja ocena disperzije srednje vrednosti je dobijena korišćenjem stratifikovanog slučajnog uzorka nad obeležjem *gender*. Srednja vrednost obeležja *writing.score* populacije iznosi $m_Y = 68.054$.

Tabela 1: Tabela rezultata obeležja *math.score*

plan	$\widehat{m_Y}$	$\widehat{D}(\widehat{m_Y})$	I_{m_Y}
PSU	66.33094	0.5717190	[64.84897, 67.81291]
S(lunch)	66.93715	0.5528149	[65.47989, 68.39442]
S(edu)	65.96367	0.5890452	[64.45941, 67.46793]
S(gender)	65.64583	0.5512826	[64.19059, 67.10107]
S(test)	65.72040	0.6379601	[64.15493, 67.28587]
S(race)	67.07476	0.4888754	[65.70436, 68.44516]

Tabela 2: Tabela rezultata obeležja *reading.score*

plan	$\widehat{m_Y}$	$\widehat{D}(\widehat{m_Y})$	I_{m_Y}
PSU	69.77338	0.5932758	[68.26373, 71.28303]
S(lunch)	69.65492	0.5162316	[68.24670, 71.06314]
S(edu)	69.24677	0.5233413	[67.82888, 70.66465]
S(gender)	68.51199	0.4953495	[67.13255, 69.89143]
S(test)	69.28594	0.5310001	[67.85771, 70.71416]
S(race)	70.42269	0.5091164	[69.02421, 71.82117]

Tabela 3: Tabela rezultata obeležja *writing.score*

plan	$\widehat{m_Y}$	$\widehat{D}(\widehat{m_Y})$	I_{m_Y}
PSU	69.30576	0.6320947	[67.74750, 70.86401]
S(lunch)	68.19831	0.5753463	[66.71164, 69.68497]
S(edu)	68.03808	0.5529261	[66.58068, 69.49549]
S(gender)	67.33151	0.5371707	[65.89502, 68.76801]
S(test)	68.32133	0.5468686	[66.87193, 69.77074]
S(race)	69.23716	0.5408679	[67.79573, 70.67859]

Konačno, u tabeli 4 se nalaze rezultati nad obeležjem *total*, koje predstavlja aritmetičku sredinu vrednosti obeležja *math.score*, *reading.score*, *writing.score*. Iz ove tabele se može uočiti da je stratifikovan slučajan uzorak imao manju ocenu disperzije srednje vrednosti obeležja u svakom slučaju od PSU, kao i da je imao najmanju ocenu disperzije kada je primenjen nad obeležjem *race/ethnicity*. Srednja vrednost obeležja *total* populacije iznosi $m_Y = 67.7706$.

5 Zaključak

Stratifikovan slučajan uzorak predstavlja bolje rešenje od prostog slučajnog uzorka bez ponavljanja prilikom izborom uzorka nad bazom podataka koja sadrži rezultate učenika srednjih škola u Sjedinjenim Američkim Državama. Ukoliko želimo da izvršimo izbor uzorka kako bismo ocenili rezultate iz čitanja i pisanja, najbolje bi bilo izabrati stratifikovan slučajan uzorak nad obeležjem *gender*, dok ukoliko želimo da izvršimo izbor uzoraka kako bismo ocenili rezultate iz matematike i ukupan rezultat, stratifikovan slučajan uzorak nad obeležjem *race/ethnicity* nam pruža najbolje

Tabela 4: Tabela rezultata obeležja *total*

plan	$\widehat{m_Y}$	$\widehat{D}(\widehat{m_Y})$	I_{m_Y}
PSU	68.47002	0.5452136	[67.02281, 69.91723]
S(lunch)	68.26346	0.4929381	[66.88738, 69.63954]
S(edu)	67.74951	0.4973219	[66.36732, 69.13169]
S(gender)	67.16311	0.5012032	[65.77554, 68.55068]
S(test)	67.77589	0.5196308	[66.36304, 69.18874]
S(race)	68.91154	0.4574636	[67.58590, 70.23718]

rezultate.

Literatura

[1] Robert V Krejcie and Daryle W Morgan. Determining sample size for research activities. *Educational and psychological measurement*, 30(3):607–610, 1970.

[2] Ljiljana Petrović. *Teorija uzoraka i planiranje eksperimenata*. Centar za izdavačku delatnost Ekonomskog fakulteta, 2007.