Analiza skupa podataka o prihodima zaposlenih u javnom sektoru države Njujork

Seminarski rad u okviru kursa
Uvod u teoriju uzoraka
Matematički fakultet

Kristina Pantelić
mi16091@alas.matf.bg.ac.rs

12. jul 2020.

Sažetak

U ovom radu čitalac će se upoznati sa bazom podataka o godišnjim prihodima zaposlenih u javnom sektoru države Njujork za period od 2011-2018. godine, metodama odabira uzorka i tehnikama ocenjivanja koje su odabrane za ovo istraživanje, kao i rezultatima koji su dobijeni njihovom primenom na skup podataka.

Sadržaj

1 Uvod

Čest slučaj je da istraživanje o obeležju od interesa nije moguće sprovesti nad čitavom populacijom. Mogući razlozi mogu biti nedostupnost čitave populacije, veliki trošak ili praktična nemogućnost ispitivanja obeležja nad svim jedinicama u populaciji. Tada se odlučujemo za uzorkovanje i ne samo to, već i za konkretan metod odabira uzorka, kao i za konkretnu tehniku ocenjivanja nepoznatih parametara. Bitna karakteristika uzorka je njegova reprezentativnost kako bi se na osnovu uzorka mogla dobiti verodostojna informacija na nivou čitave populacije tj. slika čitave populacije.

U ovom radu korišćeni metodi za odabir uzorka su metodi iz grupe verovatnosnog uzorkovanja:

1. Prost slučajan uzorak
2. Stratifikovan uzorak
3. Grupni uzorak
4. Višeetapni uzorak

Pri izboru određenih metoda odabira uzorka, korišćene su pomoćne informacije kao što su uzorkovanje sa vraćanjem i bez vraćanja sa nejednakim verovatnoćama, gde su verovatnoće odabira proporcionalne "veličini" jedinica uzorkovanja. Takođe, u slučaju ocenjivanja nepoznatih parametara, ukoliko je ustanovljena linearna koorelacija između glavnog i pomoćnog obeležja koja prolazi kroz koordinatni početak, korišćena je tehnika količničkog ocenjivanja.

Cilj istraživanja je odrediti prosečnu zaradu na nivou čitavog javnog sektora države Njojork za period od 2011-2018. godine.

1.1 Baza podataka

Baza podataka korišćena u ovom istraživanju[1] sadrži podatke o godišnjim prihodima zaposlenih u javnom sektoru države Njujork za period od 2011-2018. godine. Populacija je prirodno podeljena na četiri podsektora javnog sektora. Na kraju svake godine, u decembru, u bazu su dodavani podaci o ukupnom godišnjem prihodu zaposlenih, kao i dodatne informacije o zaposlenima, od kojih su za ovo istraživanje izdvojeni:

- ($Fiscal.Year.End.Date$) podatak o godini
- ($Group$) podatak o poziciji zaposlenog koja može biti:
  1. operaciona
  2. administrativna i sveštenička
  3. tehnička i inženjerska
  4. profesionalna
  5. menadžerska
  6. direktorska
- ($Pay.Type$) podatak o tome da li zaposleni radi puno radno vreme ili radi honorarno

Podatak o ukupnom godišnjem prihodu zaposlenog u sebi agregira informaciju o godišnjem osnovnom prihodu za konkretnu radnu poziciju zaposlenog u toku jedne godine, prihodu ostvarenim prekovremenim radom, bonusima (dostignuća koja premašuju očekivane standarde posla na toj poziciji; iznos bonus prihoda je računat na osnovu formula koja su definisana u bonusu politike za rad konkretnog radnog organa), dodatnim prihodima (isplate pojedincu za neiskorišćena sredstva za odmor ili lično vreme, provizije, podsticaji za odličnu posvećenost ili održavanje pravilnog položaja uz profesionalnu licencu) i dodantim naknadama i oštećenjima (nadoknade za ovlašćene troškove ili sve druge oblike oporezivog dohotka koji nisu uključeni u neku od pomenutih kategorija. Ovo bi moglo uključivati prilagođavanja prethodnih plaćanja kompenzacija za ispravljanje grešaka u plaćanju)[1].

Baza podataka sadrži ukupno 1.278.238 instanci, od kojih su 12.079 eliminisane iz skupa zbog nedostajućih vrednosti, što čini ukupno 1.266.159 instanci raspoloživih za istraživanje. Izdvojene su kolone od interesa pomenute u prethodnom pasusu i dodate su dve kolone koje predstavljaju redom jedinstveni identifikator instance na nivou sektora (*id*) i pripadnost instance odgovarajućem sektoru (*Sector*). Na listingu 1 prikazan je kôd kojim se vrši filtriranje podataka, a na slici 1 prikazan je izgled očišćene baze.

```
for(i in 1:length(data)){
  data_filtered[[i]] = subset(data[[i]], select = -c(Authority.Name, Last.Name,
    Middle.Initial, Has.Employees,First.Name,Title, Exempt.Indicator, Department,
    Paid.By.Another.Entity, Paid.by.State.or.Local.Government, Base.Annualized.
    Salary, Overtime.Paid, Performance.Bonus, Extra.Pay, Other.Compensation,
    Actual.Salary.Paid))

  data_filtered[[i]] = na.exclude(filter(data_filtered[[i]], (data_filtered[[i]])$
    Total.Compensation > 0))

  #Dodata nova kolona, ID instance u okviru sektora
  data_filtered[[i]]$id = seq.int(nrow(data_filtered[[i]]))

  #Dodata nova kolona, ID sektora kome instanca pripada
  data_filtered[[i]]$Sector = i
}
```

Listing 1: Primer kôda kojim se vrši filtriranje podataka

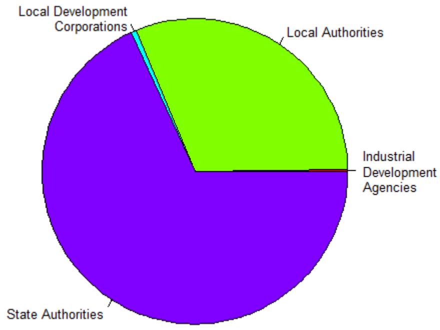
Fiscal.Year.End.Date	Group	Pay.Type	Total.Compensation	id	Sector
2018-12-31T00:00:00.000	Professional	PT	40000.00	1	1
2018-12-31T00:00:00.000	Managerial	FT	75484.00	2	1
2018-12-31T00:00:00.000	Executive	FT	120000.00	3	1
2018-12-31T00:00:00.000	Administrative and Clerical	FT	47184.00	4	1
2018-12-31T00:00:00.000	Administrative and Clerical	FT	500.00	5	1
2018-12-31T00:00:00.000	Administrative and Clerical	FT	90500.00	6	1
2018-12-31T00:00:00.000	Administrative and Clerical	FT	120961.59	7	1
2018-12-31T00:00:00.000	Managerial	FT	53469.91	8	1
2018-12-31T00:00:00.000	Administrative and Clerical	FT	98000.00	9	1
2018-12-31T00:00:00.000	Executive	FT	185500.00	10	1
2018-12-31T00:00:00.000	Administrative and Clerical	FT	22374.95	11	1
2018-12-31T00:00:00.000	Administrative and Clerical	FT	48000.00	12	1
2018-12-31T00:00:00.000	Executive	PT	48568.00	13	1
2018-12-31T00:00:00.000	Administrative and Clerical	FT	52250.00	14	1
2018-12-31T00:00:00.000	Executive	PT	52500.00	15	1

Slika 1: Izgled očišćene baze podataka

1.2 Analiza baze podataka

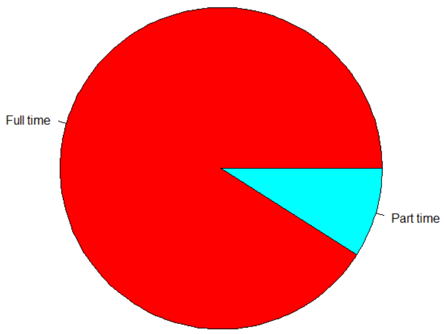
Radi boljeg uvida u sadržaj baze, u daljem tekstu je predstavljena sveobuhvatna analiza baze, analiza baze po podsektorima i različitim kriterijumima za grupu (*Group*) i tip rada (*Pay.Type*).

U skupu svih podsektora (Slika 2) postoje dva dominantna podsektora, a to su državne (*eng.* State Authorities) i lokalne vlasti (*eng.* Local Authorities). Preostala dva nedominantna podsektora čine lokalne korporacije za razvoj (*eng.* Local Developement Corporations) i agencije za industrijski razvoj (*eng.* Industrial Developement Agencies).

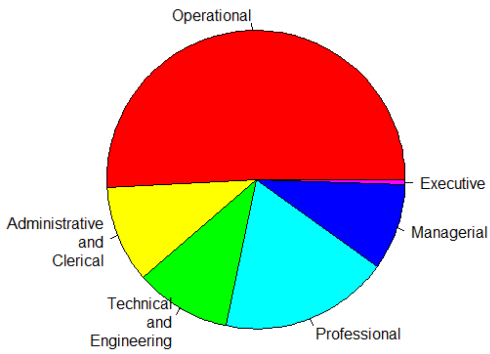


Slika 2: Podela zaposlenih po podsektorima

Na nivou čitavog javnog sektora je zastupljeniji režim rada sa punim radnim vremenom u odnosu na honoraran rad (Slika 3). Sa slike 4 se može uočiti da je najveći broj zaposlenih u grupi operacionih poslova.



Slika 3: Podela zaposlenih u zavisnosti od vrste angažmana

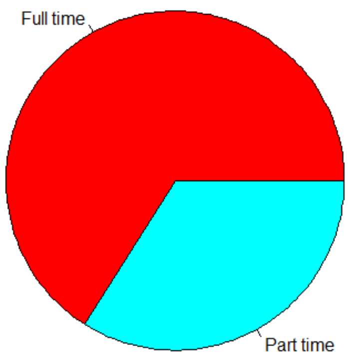


Slika 4: Podela zaposlenih u zavisnosti od radne grupe kojoj pripadaju

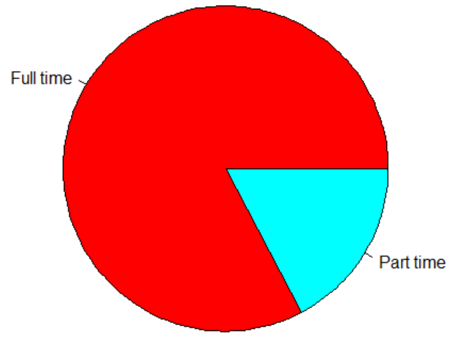
1.2.1 Analiza podsektora javnog sektora

U narednom tekstu biće prikazane dve vrste podela (u zavisnosti od vrste angažmana i radne grupe) na nivou svakog od podsektora javnog sektora.

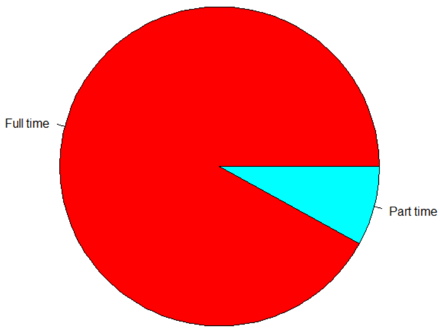
Sa slika 5, 6, 7, 8 se može uočiti da za sve podsektore važi da je većina zaposlenih u podsektoru angažovana za rad sa punim radnim vremenom, a samo određeni manji deo je angažovan za honorarni rad. Sa slika 9, 10, 11, 12 se može uočiti da su unutar svakog od podsektora zaposleni raspoređeni po radnim grupama drugačije.



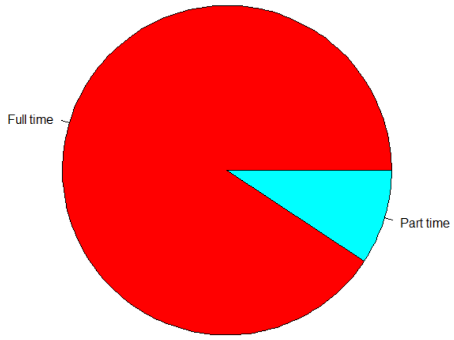
Slika 5: Podela zaposlenih u zavisnosti od vrste angažmana u okviru podsektora Agencije za industrijski razvoj



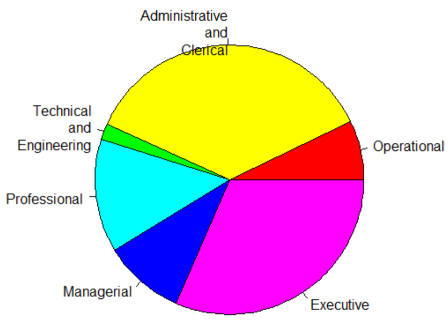
Slika 6: Podela zaposlenih u zavisnosti od vrste angažmana u okviru podsektora Lokalne korporacije za razvoj



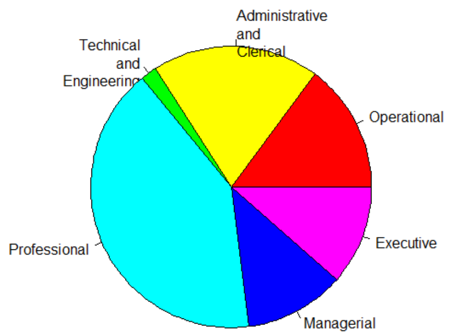
Slika 7: Podela zaposlenih u zavisnosti od vrste angažmana u okviru podsektora Lokalne vlasti



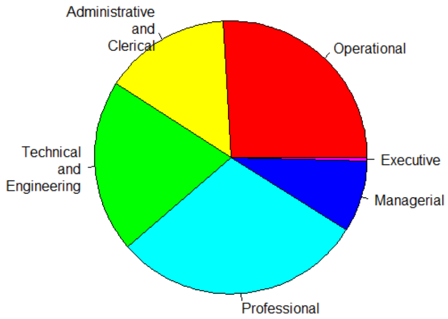
Slika 8: Podela zaposlenih u zavisnosti od vrste angažmana u okviru podsektora Državne vlasti



Slika 9: Podela zaposlenih u zavisnosti od vrste angažmana u okviru podsektora Agencije za industrijski razvoj



Slika 10: Podela zaposlenih u zavisnosti od vrste angažmana u okviru podsektora Lokalne korporacije za razvoj



Slika 11: Podela zaposlenih u zavisnosti od vrste angažmana u okviru podsektora Lokalne vlasti

Slika 12: Podela zaposlenih u zavisnosti od vrste angažmana u okviru podsektora Državne vlasti

2 Teorijske osnove i praktična primena

U ovom poglavlju su predstavljene teorijske osnove primenjenih metoda za odabir uzorka iz grupe verovatnosnog uzorkovanja:

1. Prost slučajan uzorak
2. Stratifikovan uzorak
3. Grupni uzorak
4. Višeetapni uzorak

Nakon predstavljenih teorijskih osnova svakog od metoda odabira uzorka, prikazan je i primer kôda kojim se demonstrira primena predstavljenih teorijskih osnova u praksi nad odabranom bazom podataka.

2.1 Prost slučajan uzorak

Kod prostog slučajnog uzorka, jedinica posmatranja je isto što i jedinica uzorkovanja, a odabir uzorka se vrši kroz n nezavisnih izvlačenja na slučajan način iz cele populacije. Obim populacije je označen sa N, a obim uzorka sa n. Kako uzorak može biti bez ponavljanja i sa ponavljanjem, u istraživanju su primenjena oba načina uzorkovanja.

2.1.1 Nepristrasna ocena (bez ponavljanja)

Neka je sa S označen prost slučajan uzorak bez ponavljanja. Tada se tačkasta ocena za srednju populacijsku vrednost za uzorak bez ponavljanja jedinica dobija na osnovu formule:

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

Ocena $\hat{m}_Y$ je nepristrasna tj. važi $E\hat{m}_Y = m_Y$. Ocena disperzije ove ocene data je sledećom formulom:

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

gde je $\bar{S}^2$ uzoračka disperzija.

Aproksimativni $100 \cdot (1-\alpha)\%$ dvostrani interval poverenja za nepoznatu populacijsku srednju vrednost $\hat{m}_Y$, kada je pristup zasnovan na metodu prostog slučajnog uzorkovanja bez ponavljanja jedinica, odnosno sa ponavljanjem jedinica, računa se na osnovu formula 3 i 6. Obim prostog slučajnog uzorka u istraživanju je bio $n \geq 30$, pa se na osnovu važenja Centralne granične teoreme mogao koristiti $(1 - \frac{\alpha}{2})$-kvantil standardne normalne raspodele.

$$[\hat{m}_Y - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\bar{S}^2}{n}(1-\frac{n}{N})}, \hat{m}_Y + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\bar{S}^2}{n}(1-\frac{n}{N})}] \tag{3}$$

Primer kôda kojim se dobijaju ocene u slučaju SRSWOR dat je narednim listingom.

```
n_psu = 400
salary_psu_wor = sample(lendmark_data, n_psu, replace=F)
X_salary_psu_wor = mean(salary_psu_wor)
X_salary_psu_wor

sn2_psu_wor = var(salary_psu_wor)
D_X_psu_wor = (N_pop-n_psu)*sn2_psu_wor/(N_pop*n_psu)
D_X_psu_wor
sqrt(D_X_psu_wor)

alpha = 1-0.90
z = qnorm(1-alpha/2)
I_X_psu_wor_90 = c(X_salary_psu_wor-z*sqrt(D_X_psu_wor), X_salary_psu_wor+z*sqrt(D
    _X_psu_wor))
I_X_psu_wor_90
```

Listing 2: Primer kôda kojim se dobijaju ocene u slučaju SRSWOR

2.1.2 Nepristrasna ocena (sa ponavljanjem)

Tačkasta ocena za populacijsku srednju vrednost za uzorak sa ponavljanjem jedinica se dobija na osnovu formule:

$$\hat{m}_Y = \frac{1}{n}\sum_{k=1}^{n} y_{j_k} \tag{4}$$

Ocena $\hat{m}_Y$ je takođe nepristrasna i u ovom slučaju. Ocena disperzije ove ocene date je sledećom formulom:

$$D\hat{m}_Y = \frac{\bar{S}^2}{n} \tag{5}$$

$$[\hat{m}_Y - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\bar{S}^2}{n}}, \hat{m}_Y + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\bar{S}^2}{n}}] \tag{6}$$

Primer kôda kojim se dobijaju ocene u slučaju SRSWR dat je sledećim listingom.

```
salary_psu_wr = sample(lendmark_data, n_psu, replace=T)
X_salary_psu_wr = mean(salary_psu_wr)
X_salary_psu_wr

sn2_psu_wr = var(salary_psu_wr)
D_X_psu_wr = (N_pop-n_psu)*sn2_psu_wr/(N_pop*n_psu)
D_X_psu_wr
sqrt(D_X_psu_wr)

alpha = 1-0.90
z = qnorm(1-alpha/2)
I_X_psu_wr_90 = c(X_salary_psu_wr-z*sqrt(D_X_psu_wr), X_salary_psu_wr+z*sqrt(D_X_
    psu_wr))
I_X_psu_wr_90
```

Listing 3: Primer kôda kojim se dobijaju ocene u slučaju SRSWR

2.2 Stratifikovan uzorak

Stratifikacija tj. raslojavanje je podela populacije na potpopulacije tj. slojeve koji se nazivaju stratumi. Klasifikacija svih entiteta u našoj populaciji po stratumima je vršena na osnovu nametnutih kriterijuma baze podataka. Skup podataka je inicijalno prirodno podeljen na četiri stratuma na osnovu pripadnosti zaposlenih odgovarajućem podsektoru javnog sektora. Stratumi su međusobno disjunktni tj. svaka jedinica populacije pripada tačno jednom stratumu i zadovoljavaju uslov pokrivenosti koji govori da se ne sme se pojaviti jedinica populacije koja ne pripada ni jednom stratumu [2]. Stratumi treba da imaju osobinu relativne homogenosti, ali i međusobne različitosti, što znači da vrednost obeležja treba da bude slična među jedinicama unutar stratuma, a različita među jednicama koje se nalaze u različitim stratumima [4].

Nakon izvršene stratifikacije biraju se uzorci unapred određenog obima iz svakog stratuma, a kako su stratumi međusobno nezavisni i uzorkovanje među stratumima je međusobno nezavisno. U radu je za sve stratume primenjen isti metod odabira uzorka, a to je prosto slučajno uzorkovanje.

Obimi uzoraka po stratumima određeni su proporcionalnim i Neyman-ovim optimalnim rasporedom.

Motivacija za odabir stratifikovanog uzorka u istraživanju je činjenica da su podaci u odabranoj bazi podataka inicijalno prirodno podeljeni na četiri grupacije, te hipoteza koju uvodimo u radu glasi da upotrebom stratifikovanog uzorkovanja možemo dobiti bolje ocene za nepoznatu populacijsku srednju vrednost obeležja, jer su prihodi zaposlenih unutar stratuma slični, a van stratuma se razlikuju.

2.2.1 Nepristrasna ocena (bez ponavljanja)

Neka L predstavlja broj stratuma, N_h broj jedinica u h-tom stratumu, $h = \overline{1, \mathrm{L}}$. Tada je obim populacije $N = \sum_{h=1}^{L} N_h$, obim uzorka iz h-tog stratuma $n = \sum_{h=1}^{L} n_h$, a y_{hk} je vrednost obeležja Y jedinice označene sa k koja potiče iz h-tog stratuma. Neka je S_h prost slučajan uzorak bez ponavljanja iz h-tog stratuma, $h = \overline{1, \mathrm{L}}$, nepristrasna ocena nepoznate populacijske srednje vrednosti $\hat{m}_Y$, računata je na osnovu sledeće formule:

$$\hat{m}_Y^{str} = \frac{1}{N} \sum_{h=1}^{L} N_h \cdot \bar{Y}_h \tag{7}$$

Kako je $\bar{Y}_h = \frac{1}{n_h} \sum_{k \in S_h} y_{hk}$ prethodnu formulu možemo zapisati u sledećem obliku:

$$\hat{m}_Y^{str} = \frac{1}{N} \sum_{h=1}^{L} \sum_{k \in S_h} \frac{N_h}{n_h} \cdot y_{hk} \tag{8}$$

Ocena disperzije ocene nepoznate populacijske srednje vrednosti dobija se na osnovu:

$$D\hat{m}_Y^{str} = \sum_{h=1}^{L} (\frac{N_h}{N})^2 \cdot \frac{\bar{S}_h^2}{n_h} (1 - \frac{n_h}{N_h}) \tag{9}$$

Aproksimativni $100 \cdot (1-\alpha)\%$ dvostrani interval poverenja za nepoznatu populacijsku srednju vrednost $\hat{m}_Y^{str}$ dat je sa:

$$[\hat{m}_Y^{str} - t_{(1-\frac{\alpha}{2}, n-L)} \sqrt{D\hat{m}_Y^{str}}, \hat{m}_Y^{str} + t_{(1-\frac{\alpha}{2}, n-L)} \sqrt{D\hat{m}_Y^{str}}] \tag{10}$$

Kako obim uzorka nije dovoljno veliki i ne postoji veliki broj stratuma, u formuli 10 koristi se $(1 - \frac{\alpha}{2})$-kvantil Studentove raspodele sa $(n - L)$ stepeni slobode, odnosno sa aproksimativnim brojem stepeni slobode.

Primer kôda kojim se dobijaju ocene u slučaju stratifikovanog prostog slučajnog uzorka na primeru proporcionalnog rasporeda obima uzorka dat je sledećim listingom.

```
uzorak_nh_prop = list()
for(i in 1:h){
  uzorak_nh_prop[[i]] = sample(data_filtered[[i]]$Total.Compensation, nh_prop[i], replace = F)
}
length(unlist(uzorak_nh_prop)) == n_psu #provera

tn_prop_str = c()
sn2_prop_str = c()
Di_x_prop_str = c()
for(i in 1:h){
  tn_prop_str[i] = Nh[i]*mean(uzorak_nh_prop[[i]])
  sn2_prop_str[i] = var(uzorak_nh_prop[[i]])
  Di_x_prop_str[i] = Nh[i]^2 * sn2_prop_str[i] * (1 - nh_prop[i]/Nh[i]) / nh_prop[i]
}

t_prop_str = sum(tn_prop_str)
X_prop_str = t_prop_str/N_pop
X_prop_str

D_x_prop_str = sum(Di_x_prop_str) / (N_pop^2)
sqrt(D_x_prop_str)

#Intervalna ocena
alpha = 1-0.90
z = qt(1-alpha/2, sum(n_prop) - h)
I_str_prop_90 = c(X_prop_str-z*sqrt(D_x_prop_str), X_prop_str+z*sqrt(D_x_prop_str))
I_str_prop_90
```

Listing 4: Primer kôda kojim se dobijaju ocene u slučaju stratifikovanog prostog slučajnog uzorka na primeru proporcionalnog rasporeda obima uzorka

2.2.2 Raspored obima uzorka

Za određen i fiksiran obim uzorka n, obim uzorka po stratumima određen je dvema tehnikama:

- proporcionalni raspored
- Neyman-ov optimalan raspored

Kod proporcionalnog rasporeda, broj jedinica koje se biraju u uzorak iz pojedinačnog stratuma proporcionalan je broju jedinica u tom stratumu:

$$n_h = n \cdot \frac{N_h}{N} \tag{11}$$

```
nh_prop = round(n_psu/N_pop * Nh)
if(sum(nh_prop)!=n_psu) { # ako nije = n
  while (sum(nh_prop)!=n_psu) { # ponavljamo sledece korake dok ne bude jednako n
    if(sum(nh_prop)>n_psu) {
      # ako je >n, biramo neki i smanjujemo za 1
      i = sample(1:length(nh_prop),1)
      nh_prop[i] = nh_prop[i]-1
    }
    else {
      # ako je <n, biramo neki i povecavamo za 1
      i = sample(1:length(nh_prop),1)
      nh_prop[i] = nh_prop[i]+1
    }
```

```
  }
}
sum(nh_prop) == n_psu
```

Listing 5: Primer kôda kojim se dobija obim uzorka po stratumima proporcionalnim rasporedom

Proporcionalan raspored je bolji ako su disperzije σ_h^2 koliko-toliko jednake u svim stratumima, sa stanovišta smanjenja disperzija ocena. Ako to nije slučaj, onda se pristupa Neyman-ovom optimalnom rasporedu:

$$n_h = n \cdot \frac{N_h \cdot \sigma_h}{\sum_{l=1}^{L} N_l \cdot \sigma_l} \tag{12}$$

```
nh_nejman = Nh*sqrt(si2_str)*n_psu/sum(Nh*sqrt(si2_str))
nh_nejman = round(nh_nejman)

if(sum(nh_nejman)!=n_psu) {
  while (sum(nh_nejman)!=n_psu) {
    if(sum(nh_nejman)>n_psu) {
      i = sample(1:length(nh_nejman),1)
      nh_nejman[i] = nh_nejman[i]-1
    }
    else {
      i = sample(1:length(nh_nejman),1)
      nh_nejman[i] = nh_nejman[i]+1
    }
  }
}
sum(nh_nejman) == n_psu
```

Listing 6: Primer kôda kojim se dobija obim uzorka po stratumima proporcionalnim rasporedom

S obzirom da postoje dva dominantna poddsektora u smislu broja jedinica koja im pripadaju, u istraživanju se javio slučaj da se primenom proporcionalnog i Neyman-ovog rasporeda obima uzorka dobijaju obimi uzoraka nedominantnih stratuma koji sadrže manje od dve jedinice, sa čim nije dopušteno da se nastavi istraživanje. Ovaj problem je prevaziđem malom modifikacijom kôda, a to je da se nakon određivanja obima uzoraka po stratumima, i kod proporcionalnog i kod Neyman-ovog rasporeda, uvedu dodatne provere tako da ukoliko se javi stratum sa obimom uzorka manjim od dve jedinice, tada se na slučajan način bira neki od dominantnijih stratuma, smanjuje mu se obim uzorka za jednu jedinicu, da bi se stratumu sa obimom uzorka manjim od dve jedinice broj jedinica povećao. Kôd se može videti na listingu 7.

```
while(TRUE){
  if (any(nh_prop < 2)){ # ako je neki manji
    j = which(nh_prop < 2)# pronadji koji su manji
    if(length(j) > 1){
      for(k in 1:j){
        i = sample(1:length(nh_prop),1)
        if (i \%!in\% j) { #uzmi samo one koji nemaju manje od 2 jedinice
          nh_prop[i] = nh_prop[i]-1
          nh_prop[j] = nh_prop[j]+1
        }
      }
    }
    else{
      i = sample(1:length(nh_prop),1)
      if (i != j){ #uzmi samo one koji nemaju manje od 2 jedinice
        nh_prop[i] = nh_prop[i]-1
        nh_prop[j] = nh_prop[j]+1
      }
    }
  }else{
    break
  }
```

```
}
sum(nh_prop) == n_psu
```

Listing 7: Primer kôda kojim se vrši modifikacija proporcionalnog i Neyman-ovog optimalnog rasporeda

2.3 Grupni uzorak

Kod jednoetapnog grupnog uzorkovanja važi da jedinice posmatranja nisu ujedno i jedinice uzorkovanja. Populacija se deli na primarne jedinice tj. klastere koje predstavljaju jedinice uzorkovanja i na sekundarne jedinice koje predstavljaju jedinice posmatranja.

Motivacija za grupno uzorkovanje u ovom istraživanju jeste, slično kao i kod stratifikovanog uzorkovanja, činjenica da je populacija prirodno organizovana u grupe. S obzirom da se u različitim podsektorima nalaze zaposleni istih radnih grupa, postavljamo hipotezu da bi grupni uzorak mogao dati bolje ocene nepoznate populacijske srednje vrednosti za prosečnu godišnju zaradu zaposlenih od stratifikovanog slučajnog uzorka.

2.3.1 Nepristrasna ocena (bez ponavljanja)

Sa N označavamo broj grupa, sa M_l broj sekundarnih jedinica u l-toj grupi, $l = \overline{1,\mathrm{N}}$, tada je obim populacije $M = \sum_{l=1}^{N} M_l$. Sa n označavamo obim uzorka grupa tj. primarnih jedinica, a y_{lk} je vrednost obeležja Y entiteta označenog sa k, koja potiče iz l-te grupe.

Neka je S prost slučajan uzorak bez ponavljanja. Tada je nepristrasna tačkasta ocena nepoznate populacijske srednje vrednosti:

$$\hat{m}_Y^u = \frac{N}{n \cdot M} \sum_{l \in S} \tau_l = N \cdot \bar{Y}_\tau \tag{13}$$

gde je $\bar{Y}_\tau = \frac{1}{n} \sum_{l \in S} \tau_l$.

Ocena disperzije ove ocene računa se po formuli:

$$D\hat{m}_Y^u = \frac{1}{M^2} \frac{N^2}{n} (1 - \frac{n}{N}) \bar{S}_\tau^2 \tag{14}$$

gde je $\bar{S}_\tau^2 = \frac{1}{n-1} \sum_{l \in S} (\tau_l - \bar{Y}_\tau)^2$.

Aproksimativni $100 \cdot (1 - \alpha)\%$ dvostrani interval poverenja za nepoznatu populacijsku srednju vrednost $\hat{m}_Y^u$ dat je sa:

$$[\hat{m}_Y^u - z_{1-\frac{\alpha}{2}} \sqrt{D\hat{m}_Y^u}, \hat{m}_Y^u + z_{1-\frac{\alpha}{2}} \sqrt{D\hat{m}_Y^u}] \tag{15}$$

Primer kôda kojim se dobija nepristrasna ocena grupnog prostog slučajnog uzorka dat je sledećim listingom.

```
n_group = 2
index_group_wor = sample(h, n_group, replace=F)
ti_group_psu_wor = c()
for(i in 1:h){
  ti_group_psu_wor[i] = sum(data_filtered[[i]]$Total.Compensation)
}
t_group_psu_wor = h*mean(ti_group_psu_wor[index_group_wor])
X_group_psu_wor = t_group_psu_wor / M
X_group_psu_wor

D_t_group_psu_wor = (h^2)*(1-n_group/h)*(sum((ti_group_psu_wor[index_group_wor]-
    sum(ti_group_psu_wor[index_group_wor])/n_group)^2))/(n_group*(n_group-1))
D_X_group_psu_wor = D_t_group_psu_wor / (M^2)
```

```
D_X_group_psu_wor
sqrt(D_X_group_psu_wor)

# Interval poverenja
alpha = 1-0.90
z = qnorm(1-alpha/2)
I_grwor_90 = c(X_group_psu_wor-z*sqrt(D_X_group_psu_wor), X_group_psu_wor+z*sqrt(D
    _X_group_psu_wor))
I_grwor_wor_90
```

Listing 8: Primer kôda kojim se dobija nepristrasna ocena grupnog prostog slučajnog uzorka

2.3.2 Količinska ocena (bez ponavljanja)

U podacima je ustanovljeno da su totali τ_l primarnih jedinica visoko koorelirani sa odgovarajućim "veličinama" M_l primarnih jedinica. Zahvaljujući toj osobini, izvršeno je količinsko ocenjivanje, u kome je pomoćno obeležje veličina grupe. Tačkasta ocena nepoznate populacijske srednje vrednosti data je formulom:

$$\hat{m}_Y^r = \frac{\sum_{l\in S}\tau_l}{\sum_{l\in S} M_l} \tag{16}$$

Ocena disperzije ove ocene računa se po formuli:

$$D\hat{m}_Y^r = \frac{1}{M^2}\frac{N^2}{n}(1-\frac{n}{N})\cdot\frac{1}{n-1}\sum_{l\in S}(\tau_l - b\cdot M_l)^2 \tag{17}$$

Aproksimativni $100\cdot(1-\alpha)\%$ dvostrani interval poverenja za nepoznatu populacijsku srednju vrednost $\hat{m}_Y^r$ dat je sa:

$$[\hat{m}_Y^r - z_{1-\frac{\alpha}{2}}\sqrt{D\hat{m}_Y^r}, \hat{m}_Y^r + z_{1-\frac{\alpha}{2}}\sqrt{D\hat{m}_Y^r}] \tag{18}$$

Primer kôda kojim se dobija količnička ocena grupnog prostog slučajnog uzorka dat je sledećim listingom.

```
b = cor(ti_group_psu_wor[index_group_wor], Mi[index_group_wor])
b #1
R_group_ocena = sum(ti_group_psu_wor[index_group_wor])/sum(Mi[index_group_wor])
R_group_ocena

D_t_R_group_ocena = (h^2)*(1-n_group/h)*sum((ti_group_psu_wor[index_group_wor]-R_
    group_ocena*Mi[index_group_wor])^2)/(n_group*(n_group-1))
D_R_group_ocena = D_t_R_group_ocena / (M^2)
D_R_group_ocena
sqrt(D_R_group_ocena)

# Interval poverenja
alpha = 1-0.90
z = qnorm(1-alpha/2)
I_grkol_90 = c(R_group_ocena-z*sqrt(D_R_group_ocena), R_group_ocena+z*sqrt(D_R_
    group_ocena))
I_grkol_90
```

Listing 9: Primer kôda kojim se dobija količnička ocena grupnog prostog slučajnog uzorka

2.3.3 Uzorak sa nejednakim verovatnoćama (sa i bez ponavljanja)

U ovom delu rada predstavljeno je grupno uzorkovanje sa nejednakim verovatnoćama izbora grupa tzv. uzorak grupa sa verovatnoćama proporcionalnim "veličini"grupa. U svim ocenama navedenim u narednom tekstu uzorak grupa je obima 2.

Hansen-Hurwitz-ova ocena sa ponavljanjem

Pretpostavljeno je da je slučajan uzorak R neuređeni skup kardinalnosti n, sa eventualnim ponavljanjima oznaka grupa. Za fiksirano l verovatnoća ψ_l odabira l-te grupe je data formulom:

$$\psi_l = \frac{M_l}{M} \tag{19}$$

Tada je nepristrasna tačkasta ocena nepoznate populacijske srednje vrednosti data formulom:

$$\hat{m}_Y^{\psi} = \frac{1}{n} \sum_{l \in R} \frac{\tau_l}{M_l} \tag{20}$$

Ocena disperzije ove ocene data je formulom:

$$D\hat{m}_Y^{\psi} = \frac{1}{n(n-1)} \sum_{k \in R} (m_l - \frac{\hat{\tau}_Y^{\psi}}{M})^2 \tag{21}$$

Aproksimativni $100 \cdot (1-\alpha)\%$ dvostrani interval poverenja za nepoznatu populacijsku srednju vrednost $\hat{m}_Y^{\psi}$ dat je sa:

$$[\hat{m}_Y^{\psi} - z_{1-\frac{\alpha}{2}} \sqrt{D\hat{m}_Y^{\psi}}, \hat{m}_Y^{\psi} + z_{1-\frac{\alpha}{2}} \sqrt{D\hat{m}_Y^{\psi}}] \tag{22}$$

Primer kôda kojim se dobija nepristrasna Hansen-Hurwitz-ova ocena grupnog uzorka dat je sledećim listingom.

```
n_hh = 2
pi = Mi/M
sum(pi) == 1 #provera

index_hh = sample(h, n_hh, replace=T, prob=pi)
#original podaci
ti_group_kol = c()
for(i in 1:h){
  ti_group_kol[i] = sum(data_filtered[[i]]$Total.Compensation)
}

t_hh = sum(ti_group_kol[index_hh]/pi[index_hh])/n_hh
X_hh = t_hh / M
X_hh

D_t_hh_ocena = sum((ti_group_kol[index_hh]/pi[index_hh] - t_hh)^2) / (n_hh*(n_hh
    -1))
D_X_hh_ocena = D_t_hh_ocena / (M^2)
D_X_hh_ocena
sqrt(D_X_hh_ocena)

# Interval poverenja
alpha = 1-0.90
z = qnorm(1-alpha/2)
I_grhh_90 = c(X_hh-z*sqrt(D_X_hh_ocena), X_hh+z*sqrt(D_X_hh_ocena))
I_grhh_90
```

Listing 10: Primer kôda kojim se dobija nepristrasna Hansen-Hurwitz-ova ocena grupnog uzorka

Horvitz-Thompson-ova ocena sa ponavljanjem

Neka je S' redukovani uzorak grupa sa efektivnim obimom uzorka n_D. Verovatnoća uključenja prvog reda π_l data je formulom

$$\pi_l = 1 - (1 - \psi_l)^n \tag{23}$$

gde je $\psi_l = \frac{M_l}{M}$.

Verovatnoća uključenja drugog reda π_{kl} data je formulom

$$\pi_{kl} = \pi_k + \pi_l - 1 + (1 - \psi_k - \psi_l)^n \tag{24}$$

Tada je nepristrasna tačkasta ocena nepoznate populacijske srednje vrednosti data formulom:

$$\hat{m}_Y^{\pi} = \frac{1}{M} \sum_{l \in S'} \frac{\tau_l}{\pi_l} \tag{25}$$

Ocena disperzije ove ocene data je formulom:

$$D\hat{m}_Y^{\pi} = \frac{1}{M^2} (\sum_{k \in S'} \frac{1 - \pi_k}{\pi_k^2} t_k^2 + \sum_{k \in S'} \sum_{\substack{l \in S' \\ l \neq k}}) \frac{\pi_{kl} - \pi_k \pi_l}{\pi_k \pi_l} \frac{t_k t_l}{\pi_{kl}} \tag{26}$$

Aproksimativni $100 \cdot (1 - \alpha)\%$ dvostrani interval poverenja za nepoznatu populacijsku srednju vrednost $\hat{m}_Y^{\pi}$ dat je sa:

$$[\hat{m}_Y^{\pi} - z_{1-\frac{\alpha}{2}} \sqrt{D\hat{m}_Y^{\pi}}, \hat{m}_Y^{\pi} + z_{1-\frac{\alpha}{2}} \sqrt{D\hat{m}_Y^{\pi}}] \tag{27}$$

Primer kôda kojim se dobija nepristrasna Horvitz-Thompson-ova ocena grupnog uzorka dat je sledećim listingom.

```
pii = 1-(1-pi)^n_hh
t_ht = sum(ti_group_kol[index_hh]/pii[index_hh])
X_ht = t_hh / M
X_ht
index_hh #razlicite jedinke, to je u redu

D_x_ht_ocena = sum((1-pii[index_hh])*((ti_group_kol[index_hh])^2)/((pii[index_hh])
    ^2))
for(i in index_hh) {
  for(j in index_hh) {
    if(i!=j) {
      pij = pii[i]+pii[j]-1+(1-pi[i]-pi[j])^n_hh
      D_x_ht_ocena = D_x_ht_ocena + (pij-pii[i]*pii[j])*(ti_group_kol[i]*ti_group_
    kol[j])/(pii[i]*pii[j]*pij)
    }
  }
}
D_x_ht_ocena = D_x_ht_ocena/(M^2)
D_x_ht_ocena
sqrt(D_x_ht_ocena)

# Interval poverenja
alpha = 1-0.90
z = qnorm(1-alpha/2)
I_grht_90 = c(X_ht-z*sqrt(D_x_ht_ocena), X_ht+z*sqrt(D_x_ht_ocena))
I_grht_90
```

Listing 11: Primer kôda kojim se dobija nepristrasna Horvitz-Thompson-ova ocena grupnog uzorka

Sen-Yates-Grundy-jeva ocena bez ponavljanja

Kako za uzorak bez ponavljanja ne važi u opštem slučaju formula (23), verovatnoće uključenja prvog reda određujemo formirajući sve moguće uzorke željenog obima, a zatim sabiramo verovatnoće uzoraka ukoliko se jedinica nalazi u uzorku.

Neka je S slučajan uzorak bez ponavljanja obima n. Nepristrasna tačkasta ocena nepoznate populacijske srednje vrednosti se računa kao i u prethodnom slučaju, po formuli (25). Tačkasta ocena disperzije ocene nepoznate populacijske srednje vrednosti se računa na osnovu sledeće formule:

$$D\hat{m}_Y^{\pi} = \frac{1}{2 \cdot M^2} \sum_{k \in S} \sum_{\substack{l \in S \\ l \neq k}} \frac{\pi_k \pi_l - \pi_{kl}}{\pi_{kl}} (\frac{t_k}{\pi_k} - \frac{t_l}{\pi_l})^2 \tag{28}$$

Aproksimativni $100 \cdot (1 - \alpha)\%$ dvostrani interval poverenja za nepoznatu populacijsku srednju vrednost $\hat{m}_Y^{\pi}$ dat je formulom 27.

Primer kôda kojim se dobija nepristrasna Sen-Yates-Grundy-jeva ocena grupnog uzorka dat je sledećim listingom.

```
uzorci = list()
#obelezje_na_uzorku = list()
i = 1
for(j in 1:h) {
  for(k in 1:h) {
    if(j<k) {
      uzorci[[i]] = c(j,k)
      i = i+1
    }
  }
}
length(uzorci) ==choose(4,2) #provera

vca_uzorka = function(i,j) {
  (Mi[i]/M)*(Mi[j]/(M-Mi[i])) + (Mi[j]/M)*(Mi[i]/(M-Mi[j]))
}

p_uzorka = c()
for(i in 1:length(uzorci)) {
  p_uzorka[i] = vca_uzorka(uzorci[[i]][1], uzorci[[i]][2])
}
sum(p_uzorka) # provera

#za uzorak bez ponavljanja ne vazi u opstem slucaju formula
pii2 = rep(0,h)
for(i in 1:h) {
  for(j in 1:length(uzorci)) {
    if(i \%in\% uzorci[[j]]) {
      pii2[i] = pii2[i]+p_uzorka[j]
    }
  }
}
sum(pii2) == n_hh #provera

index_syg = sample(h, n_hh, replace=F, prob=pi)
t_syg = sum(ti_group_kol[index_syg]/pii2[index_syg])
X_syg = t_syg / M
X_syg

D_t_ht_ocena_syg = 0
for (i in index_syg) {
  for(j in index_syg) {
    if(i<j) {
      pij = pii2[i]+pii2[j]-1+(1-pi[i]-pi[j])^n_hh
      D_t_ht_ocena_syg = D_t_ht_ocena_syg + (pii2[i]*pii2[j]-pij)*((ti_group_kol[i
    ]/pii2[i]-ti_group_kol[j]/pii2[j])^2)/pij
    }
  }
}
D_x_ht_ocena_syg = D_t_ht_ocena_syg / ((M^2)*2)
D_x_ht_ocena_syg
sqrt(D_x_ht_ocena_syg)

# Interval poverenja
alpha = 1-0.90
z = qnorm(1-alpha/2)
I_grht_90 = c(X_syg-z*sqrt(D_x_ht_ocena_syg), X_syg+z*sqrt(D_x_ht_ocena_syg))
I_grht_90
```

Listing 12: Primer kôda kojim se dobija nepristrasna Sen-Yates-Grundy-jeva ocena grupnog uzorka

2.4 Višeetapni uzorak

Kod višeetapnog uzorkovanja podela populacije je izvršena na primarne, sekundarne, tercijarne jedinice itd. Jedinice posmatranja su podele, a jedinice uzorkovanja su elementi odgovarajuće podele. U ovom radu izvršeno je dvoetapno grupno uzorkovanje, koje se sprovodi tako što se odabere određeni broj primarnih jedinica, a zatim se bira uzorak sekundarnih jedinica iz svake odabrane primarne jedinice.

Motivacija za primenu dvoetapnog uzorkovanja u istraživanju je to što su klasteri, koji su prirodno nametnuti u bazi podataka, veoma različitih dimenzija. Velika razlika u broju jedinica među klasterima može da predstavlja dominaciju određenog klastera u oceni nepoznate populacijske srednje vrednosti, te je ideja da se umesto čitavog klastera u uzorak uzme samo određeni broj jedinica iz odabranih primarnih jedinica. Obim sekundarnih jedinica po klasterima je određen proporcionalno broju jedinica u klasteru.

2.4.1 Nepristrasna ocena (bez ponavljanja)

Neka je N broj grupa, M_l je broj sekundarnih jedinica u l-toj grupi, $l = \overline{1, N}$, obim populacije je $M = \sum_{l=1}^{N} M_l$. Neka je n obim uzorka S grupa tj. primarnih jedinica, n_l je obim uzorka S_l entiteta tj. sekundarnih jedinica iz l-te grupe, $l = \overline{1, N}$, a y_{lk} je vrednost obeležja Y entiteta označenog sa k koji potiče iz l-te grupe.

Tačkasta ocena nepoznate populacijske srednje vrednosti data je formulom:

$$\hat{m}_Y^u = \frac{1}{M} \frac{N}{n} \sum_{l \in S} \hat{\tau}_l = \frac{1}{M} \sum_{l \in S} \sum_{k \in S_l} \frac{N}{n} \cdot \frac{M_l}{n_l} \cdot y_{lk} \tag{29}$$

gde je $\hat{\tau}_l = \frac{M_l}{n_l} \sum_{k \in S_l} y_{lk} = M_l \cdot \bar{Y}_l$.

Ocena disperzije ove ocene data je formulom:

$$D\hat{m}_Y^u = \frac{1}{M^2} (\frac{N^2}{n} (1 - \frac{n}{N}) \bar{S}_\tau^2 + \frac{N}{n} \sum_{l \in S} \frac{M_l^2}{n_l} (1 - \frac{n_l}{M_l}) \bar{S}^2) \tag{30}$$

gde su

$$\bar{S}_\tau^2 = \frac{1}{n-1} \sum_{l \in S} (\hat{\tau}_l - \bar{Y}_\tau)^2$$

$$\bar{Y}_\tau = \frac{\hat{\tau}_Y^u}{N}$$

i

$$\bar{S}_l^2 = \frac{1}{n_l - 1} \sum_{k \in S_l} (y_{lk} - \bar{Y}_l)^2$$

Aproksimativni $100 \cdot (1 - \alpha)\%$ dvostrani interval poverenja za nepoznatu populacijsku srednju vrednost $\hat{m}_Y^u$ dat je sa:

$$[\hat{m}_Y^u - t_{(1-\frac{\alpha}{2}, n-L)} \sqrt{D\hat{m}_Y^u}, \hat{m}_Y^u + t_{(1-\frac{\alpha}{2}, n-L)} \sqrt{D\hat{m}_Y^u}] \tag{31}$$

Kako obim uzorka nije dovoljno veliki i ne postoji veliki broj grupa, a u nekim grupama može biti i manji broj sekundarnih jedinica u uzorku koje ne bi zadovoljavale uslov $n \geq 30$, u formuli 31 se koristi $(1 - \frac{\alpha}{2})$-kvantil Studentove raspodele sa $(n - L)$ stepeni slobode, odnosno sa aproksimativnim brojem stepeni slobode.

Primer kôda kojim se dobija nepristrasna ocena dvoetapnog uzorka dat je sledećim listingom.

```
sekundarne = list()
Mi_mgroup=c()
for(i in 1:n_group){
  # biram proporcionalno obimu uzorka...
  sekundarne[[i]] = sample(data_filtered[[index_group_wor[i]]]$Total.Compensation,
      nh_prop[index_group_wor[i]], replace=F)
  Mi_mgroup[i] = Mi[index_group_wor[i]]
}
ti_mgroup_psu_wor = c() #uzorkovane
for(i in 1:n_group){
```

```
  ti_mgroup_psu_wor[i] = Mi_mgroup[i]*mean(sekundarne[[i]])
}
t_mgroup_psu_wor = h*mean(ti_mgroup_psu_wor)
X_mgroup_psu_wor = t_mgroup_psu_wor / M
X_mgroup_psu_wor

s_t2 = sum((ti_mgroup_psu_wor-t_mgroup_psu_wor/h)^2)/(n_group-1)
s_i2 = c()
for(i in 1:n_group) {
  s_i2[i] = var(sekundarne[[i]])
}

D_t_mgroup_psu_wor = h^2*(1-n_group/h)*s_t2/(n_group) + h*sum(Mi_mgroup^2*(1-nh_
    prop[index_group_wor]/Mi_mgroup)*s_i2/nh_prop[index_group_wor])/n_group
D_X_mgroup_psu_wor = D_t_group_psu_wor / (M^2)
D_X_mgroup_psu_wor
sqrt(D_X_mgroup_psu_wor)

#Intervalna ocena
alpha = 1-0.90
z = qt(1-alpha/2, sum(n_prop) - h)
I_multiwor_90 = c(X_mgroup_psu_wor-z*sqrt(D_X_mgroup_psu_wor), X_mgroup_psu_wor+z*
    sqrt(D_X_mgroup_psu_wor))
I_multiwor_90
```

Listing 13: Primer kôda kojim se dobija nepristrasna ocena dvoetapnog uzorka

2.4.2 Količinska ocena (bez ponavljanja)

Količinskom ocenom dobijena je tačkasta ocena nepoznate populacijske srednje vrednosti, jer postoji visoka koorelacija između totala primarnih jedinica i odgovarajućih veličina primarnih jedinica. Ova ocena nije nepristrasna.

$$\hat{m}_Y^r = \frac{\sum_{l\in S}\hat{\tau}_l}{\sum_{l\in S}M_l} \tag{32}$$

Ocena disperzije ove ocene računa se po formuli:

$$D\hat{m}_Y^r = \frac{1}{M^2}\left(\frac{N^2}{n}(1-\frac{n}{N})\frac{1}{n-1}\sum_{l\in S}(\hat{\tau}_l-\hat{b}\cdot M_l)^2 + \frac{N}{n}\sum_{l\in S}\frac{M_l^2}{n_l}(1-\frac{n_l}{M_l})\bar{S}^2\right) \tag{33}$$

Aproksimativni $100\cdot(1-\alpha)\%$ dvostrani interval poverenja za nepoznatu populacijsku srednju vrednost $\hat{m}_Y^r$ dat je sa:

$$[\hat{m}_Y^r - t_{(1-\frac{\alpha}{2},n-L)}\sqrt{D\hat{m}_Y^r}, \hat{m}_Y^r + t_{(1-\frac{\alpha}{2},n-L)}\sqrt{D\hat{m}_Y^r}] \tag{34}$$

Kako obim uzorka nije dovoljno veliki i ne postoji veliki broj grupa, a u nekim grupama može biti i manji broj sekundarnih jedinica u uzorku koje ne bi zadovoljavale uslov $n \geq 30$, u formuli 31 se koristi $(1-\frac{\alpha}{2})$-kvantil Studentove raspodele sa $(n-L)$ stepeni slobode, odnosno sa aproksimativnim brojem stepeni slobode.

Primer kôda kojim se dobija količinska ocena dvoetapnog uzorka dat je sledećim listingom.

```
b_m = cor(ti_mgroup_psu_wor, Mi_mgroup)
b_m #1
R_mgroup_ocena = sum(ti_mgroup_psu_wor)/sum(Mi_mgroup)
R_mgroup_ocena

s_t2_R = sum((ti_mgroup_psu_wor-R_mgroup_ocena*Mi_mgroup)^2)/(n_group-1)
D_t_R_mgroup_psu_wor = h^2*(1-n_group/h)*s_t2_R/n_group + h*sum(Mi_mgroup^2*(1-nh_
    prop[index_group_wor]/Mi_mgroup)*s_i2/nh_prop[index_group_wor])/n_group
D_X_R_mgroup_psu_wor = D_t_R_mgroup_psu_wor / (M^2)
D_X_R_mgroup_psu_wor
sqrt(D_X_R_mgroup_psu_wor)

```

```
#Intervalna ocena
alpha = 1-0.90
z = qt(1-alpha/2, sum(n_prop) - h)
I_multikol_90 = c(R_mgroup_ocena-z*sqrt(D_X_R_mgroup_psu_wor), R_mgroup_ocena+z*
    sqrt(D_X_R_mgroup_psu_wor))
I_multikol_90
```

Listing 14: Primer kôda kojim se dobija količinska ocena dvoetapnog uzorka

3 Rezultati i diskusija

Izbor veličine prostog slučajnog uzorka je dobijen na osnovu podataka o obimu populacije iz literature [3]. U istraživanju su korišćena četiri metoda odabira uzorka čije su teorijske osnove prikazane u prethodnom poglavlju, a čiji će rezultati primene nad podacima biti prikazani u tekućem poglavlju. Treba imati u vidu da prava populacijska srednja vrednost posmatranog obeležja iznosi 68128.81 dolara.

3.1 Prosto slučajno uzorkovanje

U tabeli 1 su dati rezultati za prost slučajan uzorak sa ponavljanjem (**SRSWR**) i bez ponavljanja (**SRSWOR**), njihove ocene disperzija i aproksimativni 90% dvostrani intervali poverenja za nepoznatu populacijsku srednju vrednost.

Tabela 1: Rezultati za ocenjenu nepoznatu populacijsku srednju vrednost i ocenjene standardne greške za prost slučajan uzorak

plan	$\hat{m}_Y$	$\sqrt{\hat{D}\hat{m}_Y}$	I_{m_Y}
SRSWOR	66824.04	2201.254	$[63203.30, 70444.78]$
SRSWR	68753.89	2038.202	$[65401.34, 72106.43]$

Na osnovu ocena standardnih grešaka, može se uočiti da je prosto slučajno uzorkovanje sa ponavljanjem jedinica u uzorku dalo bolje rezultate nego što je to dalo prosto slučajno uzorkovanje bez ponavljanja jedinica.

3.2 Stratifikovano uzorkovanje

U tabeli 2 su dati rezultati za stratifikovan prost slučajan uzorak bez ponavljanja, dobijen primenom proporcionalnog (**StrProp**) i Neyman-ovog optimalnog rasporeda (**StrNeyman**) za odabir veličine uzorka; date su njihove ocene disperzija i aproksimativni 90% dvostrani intervali poverenja za nepoznatu populacijsku srednju vrednost.

Tabela 2: Rezultati za ocenjenu nepoznatu populacijsku srednju vrednost i ocenjene standardne greške za stratifikovano uzorkovanje

plan	$\hat{m}_Y$	$\sqrt{\hat{D}\hat{m}_Y}$	I_{m_Y}
StrProp	69311.91	2026.173	$[63983.93, 69887.41]$
StrNeyman	69520.91	2013.566	$[63778.72, 70434.86]$

Poređenjem vrednosti ocena standardnih grešaka, uočavamo da su greške manje kod stratifikovanog uzorka nego kod prostog slučajnog uzorkovanja sa i bez ponavljanja. Dodatno, možemo uočiti da se Neyman-ov optimalan raspored obima uzorka pokazao kao bolji izbor na konkretnim podacima.

3.3 Grupno uzorkovanje

U tabeli 3 su dati rezultati za grupni uzorak gde je ocena formirana na osnovu

- prostog slučajnog uzorkovanja bez ponavljanja primarnih jedinica u uzorku (**GrSRSWOR**)
- količničke ocene, gde je kao pomoćno obeležje korišćen podatak o obimu grupa (**GrKolic**)
- Hunsen-Hurwitz-ove ocene za uzorak sa ponavljanjem jedinica u uzorku (**GrHH**)
- Horvitz-Thompson-ove ocene za uzorak sa ponavljanjem jedinica u uzorku (**GrHT**)
- Sen-Yates-Grundy-jeve ocene za uzorak bez ponavljanja jedinica u uzorku (**GrSYG**)

Date su njihove ocene disperzija i aproksimativni 90% dvostrani intervali poverenja za nepoznatu populacijsku srednju vrednost.

Tabela 3: Rezultati za ocenjenu nepoznatu populacijsku srednju vrednost i ocenjene standardne greške za grupno uzorkovanje

plan	$\hat{m}_Y$	$\sqrt{\hat{D}\hat{m}_Y}$	I_{m_Y}
GrSRSWOR	98845.46	69692.35	$[63156.38, 207527.11]$
GrKolic	68208.23	8136.872	$[54824.26, 81592.19]$
GrHH	65690.62	6739.768	$[54604.69, 76776.55]$
GrHT	65690.62	21184.71	$[30844.87, 100536.37]$
GrSYG	68387.54	188.4797	$[68077.52, 68697.57]$

U slučaju grupnog uzorkovanja, analizirajući dobijene rezultate predstavljene u tabeli 3, zaključujemo da iako smo postavili hipotezu da bi grupni uzorak mogao dati bolje rezultate u smislu smanjenja ocena disperzije ocena, uočavamo da to nije slučaj sa svim vrstama grupnog uzorkovanja. Grupni prost slučajan uzorak bez ponavljanja primarnih jedinica u uzorku je dao rezultat ocene nepoznate populacijske srednje vrednosti koji daleko premašuje pravu vrednost, na šta ukazuje i vrednost ocene standardne greške. Suprotno od toga, već upotrebom količničkog ocenjivanja dobijaju se znatno bolje ocene, kao i ocene standardne greške, koja je znatno bolja nego u prethodnom slučaju, ali lošija od dva prethodno viđena metoda odabira uzorkovanja (prostog slučajnog uzorka i stratifikovanog uzorka). Ono što se od svih do sada pomenutih ocena i prikazanih rezultata izdvojilo je Sen-Yates-Grundy-jeva ocena za uzorak bez ponavljanja jedinica u uzorku, čija je ocena standardne greške oko 188 dolara.

3.4 Višeetapno uzorkovanje

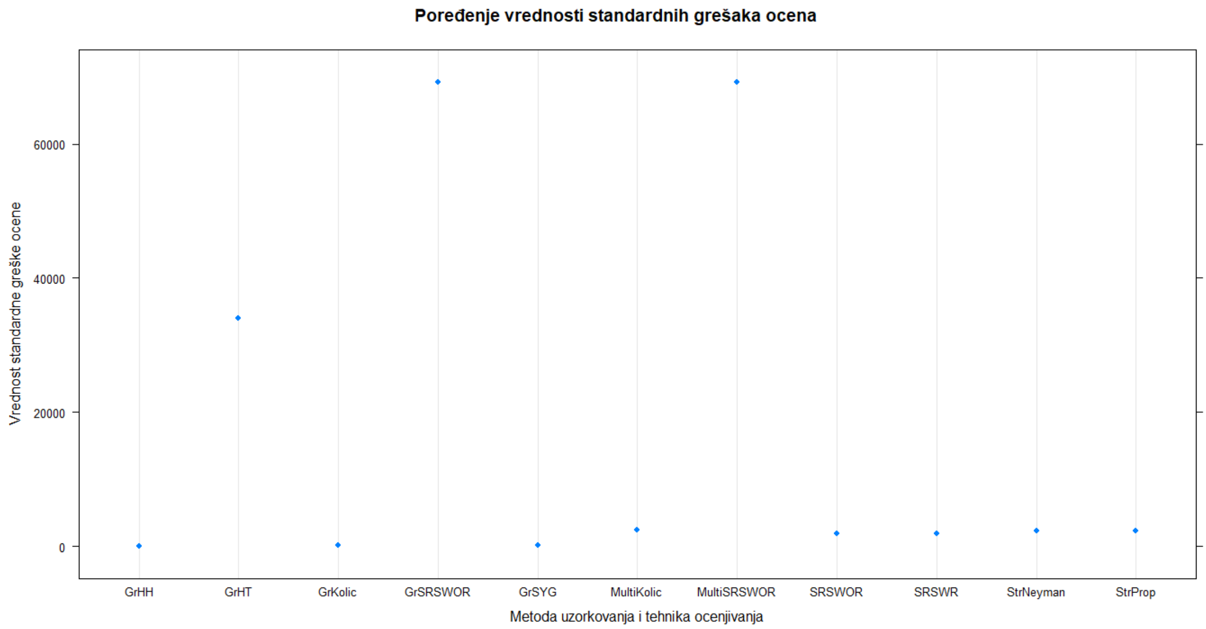
U tabeli 4 su dati rezultati za dvoetapni uzorak u kome se primarne i sekundarne jedinice u uzorak biraju metodom prostog slučajnog uzorkovanja bez ponavljanja jedinica u uzorku (**MultiSRSWOR**), a dobijena je i količnička ocena zahvaljujući visokoj koorelaciji glavnog i pomoćnog obeležja (**MultiKolic**).

Tabela 4: Rezultati za ocenjenu nepoznatu populacijsku srednju vrednost i ocenjene standardne greške za dvoetapno uzorkovanje

plan	$\hat{m}_Y$	$\sqrt{\hat{D}\hat{m}_Y}$	I_{m_Y}
MultiSRSWOR	129983.3	43885.58	$[61828.84, 206885.18]$
MultiKolic	67064.07	7333.996	$[47393.84, 88030.06]$

Kao i u slučaju grupnog prostog slučajnog uzorka bez ponavljanja, iz 4 možemo uočiti da ocene dobijene dvoetapnim prostim sučajnim uzorkovanjem nisu dobre, dok je količnička ocena znatno bolja, ali ne i bolja od svih do sada predstavljenih ocena.

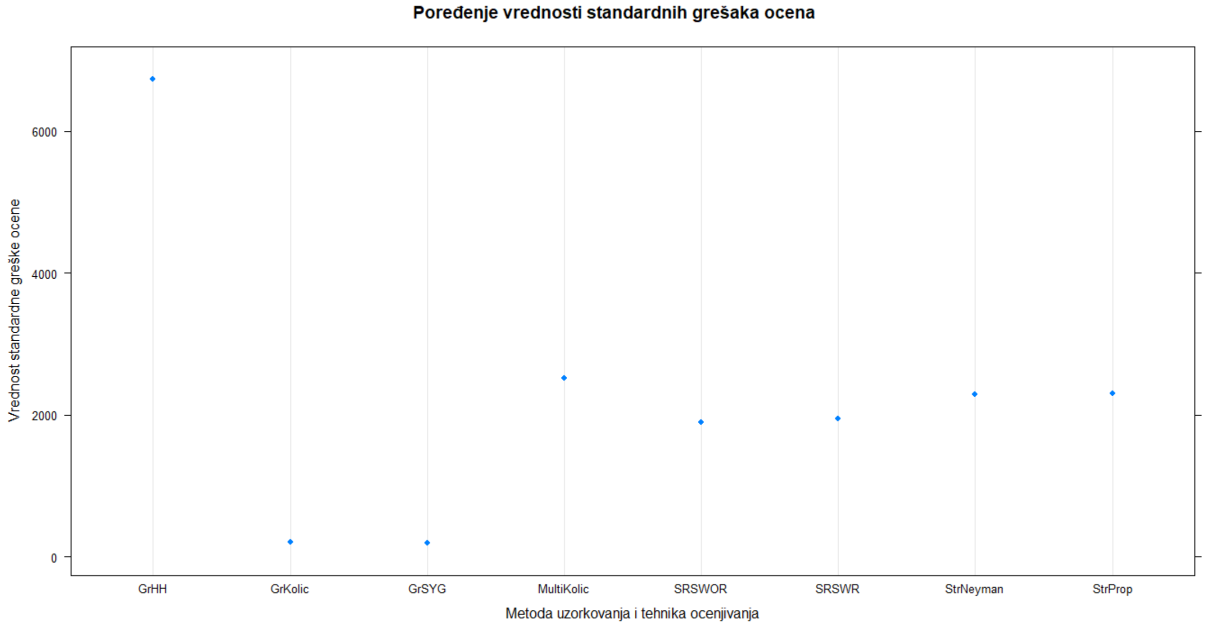
Vizuelni prikaz svih predstavljenih ocena može se videti na slici 13. Na x-osi predstavljeni su metodi odabira uzorka zajedno sa tehnikom ocenjivanja, a na y-osi je predstavljena ocena standardne greške. Možemo uočiti veliko odstupanje grupnog i dvoetapnog prostog slučajnog uzorka bez ponavljanja u odnosu na druge ocene.



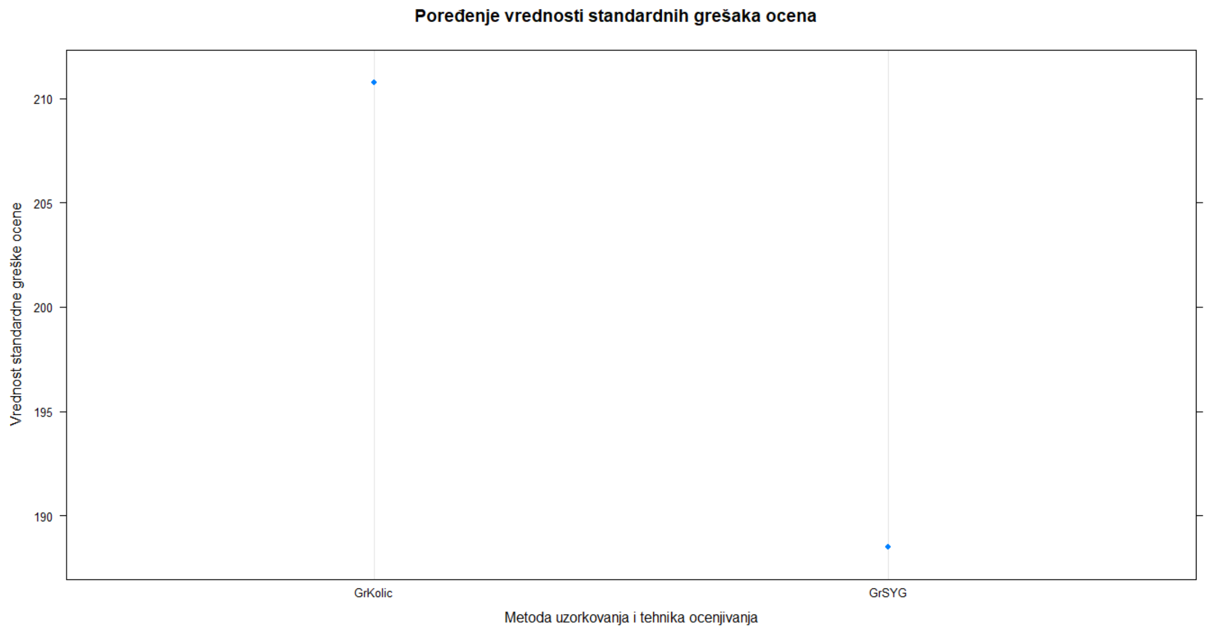
Slika 13: Poređenje ocena standardnih grešaka različitih metoda odabira uzorka i tehnika ocenjivanja 1

Radi dublje analize, odbacićemo razmatranje grupnog (**GrSRSWOR**), dvoetapnog slučajnog uzorka(**MultiSRSWOR**) i Horvitz-Thomson-ove ocene grupnog uzorka (**GrHT**), zbog velikog odstupanja u odnosu na ostale ocene i u odnosu na pravu vrednost posmatranog obeležja. Na slici 14 može se videti poređenje ostalih dobijenih ocena standardne greške.

Sa slike 14 uočavamo da najmanju ocenu standardne greške imaju Sen-Yates-Grundy-jeva ocena za prost slučajan grupni uzorak bez ponavljanja i količnička ocena za prost slučajan grupni uzorak bez ponavljanja. Kako su, posmatrajući sliku 14, ove dve vrednosti veoma bliske, još jedan detaljniji prikaz ove dve ocene se može videti na slici 15.



Slika 14: Poređenje ocena standardnih grešaka različitih metoda odabira uzorka i tehnika ocenjivanja 2



Slika 15: Poređenje ocena standardnih grešaka različitih metoda odabira uzorka i tehnika ocenjivanja 3

4 Zaključak

Baza podataka sadrži prirodnu podelu javnog sektora američke države Njujork na četiri nezavisna podsektora. Ovi prirodno nametnuti podsektori u ovom istraživanju predstavlja-ju stratume kod stratifikovanog uzorka, odnosno grupe kod grupnog uzorka. Na početku istraživanja je postavljena hipoteza da se raslojavanjem populacije mogu dobiti bolje ocene za nepoznatu populacijsku srednju vrednost godišnjih prihoda zaposlenih u javnom sektoru države Njujork. Na osnovu teorijskih osnova datim u poglavlju 2 u programskom jeziku R su izvršena izračunavanja i dobijene su ocene nepoznatog parametra, ocene disperzija ocena, kao i 90% aproksimativni intervali poverenja nepoznatog parametra. U poglavlju 3 dati su rezultati istraživanja i na osnovu datih rezultata zaključujemo da se od svih predstavljenih metoda odabira uzorka i tehnika ocenjivanja istakla Sen-Yates-Grundy-jeva ocena nejednake verovatnoće izbora grupa. Ovim rezultatom je potvrđena hipoteza postavljena na početku istraživanja, a to je da se zbog prirodnog raslojavanja populacije na podsektore i uzimanjem dominantnih podsektora u uzorak dobijaju bolje ocene u smislu smanjenja disperzije ocene. Hipoteza za višeetapno uzorkovanje je nakon dobijenih rezultata istraživanja odbačena, jer se pokazalo da se dodatnim uzorkovanjem sekundarnih jedinica u okviru podgrupa ne dobijaju značajno bolji rezultati.

Literatura

[1] Baza podataka o prihodima u javnom sektoru države Njujork za period od 2011-2018. godine sa opisom baze. on-line at: https://www.kaggle.com/new-york-state/nys-salary-information-for-the-public-sector.

[2] doc. dr Lenka Glavaš. Materijali za kurs Uvod u teoriju uzoraka profesorke Lenke Glavaš, 2020. on-line at: http://www.matf.bg.ac.rs/p/lenka-zivadinovic/pocetna/.

[3] Robert V. Krejcie and Daryle W. Morgan. Determing Sample Size For Research Activities. *Educational and Psychological measurement*, 38(30):607–610, 1970.

[4] Mirjana Veljović. Materijali za kurs Uvod u teoriju uzoraka asistentkinje Mirjane Veljović, 2020. on-line at: http://www.matf.bg.ac.rs/p/-mirjana-veljovic.