Analiza podataka o uspehu učenika na završnom ispitu iz Matematike u dve škole u Portugaliji

Seminarski rad u okviru kursa
Uvod u teoriju uzorka
Matematički fakultet

Dušica Golubović 119/2016

22.Jun 2020.

Sadržaj

1 Uvod

Cilj ovog istraživanja je analiza rezultata na završnom ispitu iz Matematike. Naime, poznato nam je da raspodela poena/ocena na svakom ispitu teži da ima normalnu raspodelu kao i da će prosečna ocena/prosečni poeni na testu biti polovina poena. Videćemo da li se to odnosi i na ove dve škole.

1.1 Analiza baze podataka

U ovom delu ćemo se upoznati sa samom bazom. Baza ima ukupno 395 entiteta i 33 atributa (kolona). Navešćemo atribute i njihova objašnjenja:

- **school**: binarni kategorički atribut koji ima dve vrednosti – 'GP' (Gabriel Pereira) i 'MS' (Mousinho de Silveira) koje predstavljaju dve škole u Portugaliji.
- **sex**: binarni atribut predstavlja pol učenika
- **age**: kvantitativno diskretno obeležje koje je u intervalu [15,22]
- **address**: binarni atribut koji predstavlja adresu učenika – 'U' (gradsko naselje) ili 'R' (seosko naselje)
- **famsize**: binarno obeležje koje predstavlja veličinu porodice – 'LE3' (manje ili jednako 3 člana), 'GT3' (više od 3 člana)
- **Pstatus**: binarni atribut koji opisuje bračni status roditelja – 'T' (žive zajedno), 'A' (razvedeni su)
- **Medu/Fedu**: ordinalno obeležje koje predstavlja obrazovanje majke-/oca učenika
- **Mjob/Fjob**: posao majke/oca učenika, nominalno obeležje
- **reason**: razlog za odabir škole, nominalno obeležje – blizu mesta boravka, reputacije škole, zbog nekih kurseva ili drugi razlog
- **guarding**: nominalno obeležje koje predstavlja staratelja učenika
- **traveltime**: vreme potrebno za dolazak do škole
- **studytime**: nedeljno vremene koje je učenik proveo učeći;

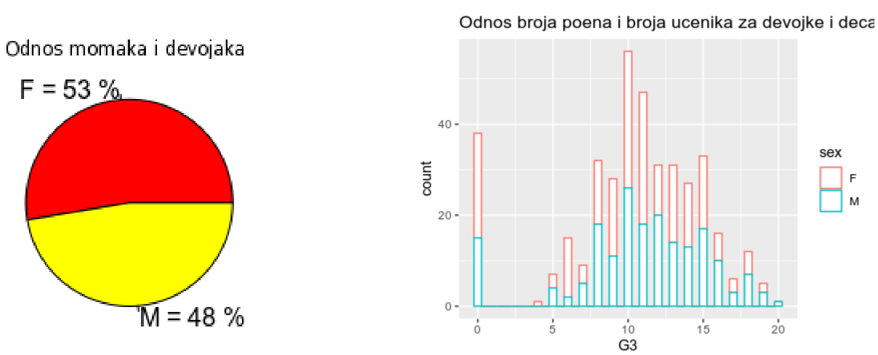
- 1 = $<$2 sata
- 2 = 2 do 5 sati
- 3 = 5 do 10 sati
- 4 = $>$10 sati

- **failures**: broj predmeta koji su padali – numeričko obeležje
- **schoolsup**: binarno obeležje koje predstavlja da li je učeniku bila potrebna dodatna nastava
- **famsup**: binarno obeležje koje predstavlja da li je učeniku bila potrebna podrška od porodice u nastavi
- **paid**: binarno obeležje koje predstavlja da li je učenik plaćao dodatne časove iz ovog predmeta
- **activities**: binarno obeležje koje predstavlja da li je učenik pohađao neke vannastavne aktivnosti
- **nursery**: binarno obeležje, da li je učenik išao u obdanište
- **higher**: binarno obeležje, da li učenik želi da ide na fakultet
- **internet**: binarno obeležje, da li učenik ima pristup internetu
- **romantic**: da li učenik ima partnera
- **famrel**: ordinalno obeležje koje predstavlja kvalitet odnosa u porodici – 1(veoma loše) - 5 (veoma visoko)
- **freetime**: ordinalno obeležje koje predstavlja količinu slobodnog vremena posle škole – 1(veoma malo) - 5 (dosta)
- **goout**: ordinalno obeležje koje predstavlja koliko učenik izlazi sa prijateljima – 1(veoma malo) - 5 (dosta)
- **Dalc/Walc**: ordinalno obeležje koje predstavlja konzumiranje alkohola tokom radnih dana/vikendom – 1(veoma mali unos) - 5 (veoma veliki unos)
- **health**: ordinalno obeležje koje predstavlja opšte zdravstveno stanje učenika – 1(veoma loše) - 5 (veoma dobro)

- **absences**: diskretno obeležje koje predstavlja broj izostanaka, u intervalu [0,93]
- **G1/G2**: ocena iz matematike u prvom/drugom polugodištu, diskretna vrednost u intervalu [0,20]
- **G3**: finalna ocena iz matematike,diskretna vrednost u intervalu [0,20]

2 Analiza i vizuelizacija podataka

Da bismo bolje razumeli samu bazu potrebno je da izvršimo detaljniju analizu podataka.



(a) Pitasti dijagram odnosa momaka i devojaka u školama

(b) Histogram broja poena za momke i devojke

Slika 1: Vizuelizacija obeležja Sex

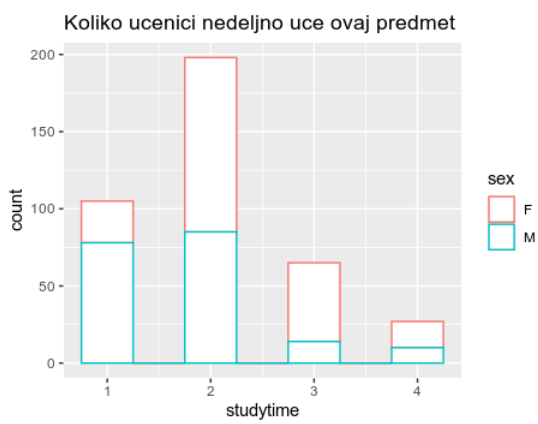
Sa slike 1 vidimo da je u našoj bazi odnos momaka i devojaka približno isti. Dalje u nastavku je data raspodela ocena za momke i devojke. Vidimo sa slike da su i momci i devojke ujednačeni kad je završna ocena u pitanju. Najveći broj momaka i devojaka ima prosečnu ocenu, dok mali broj ima veće ocene. Kada gledamo prosečnu vrednost za obeležje G3 koje predstavlja finalnu ocenu vidimo da za momke ona iznosi $m_{G3}^{M} = 10.91444$, dok je za devojke ova vrednost nešto manja i iznosi $m_{G3}^{F} = 9.966$. Kod za računanje ovih vrednosti je dat u nastavku.

```
devojke <-subset(studenti, sex == 'F')
# populacijska srednja vrednost za devojke
m_g3_devojke<-mean(devojke$G3)
```

```
momci <-subset(studenti,sex=='M')
m_g3_momci<-mean(momci$G3)
```

Još jedno obeležje za koje ćemo analizirati vrednosti dobijene za momke i devojke je obeležje studytime koje predstavlja broj nedeljnih sati koje učenici provedu učeći ovaj predmet. Njegove oznake su objašnjene u Uvodu 1.1 Kao što se vidi sa slike devojke više svog vremena posvećuju ovom predmetu. Kod za ovu sliku je dat u nastavku.

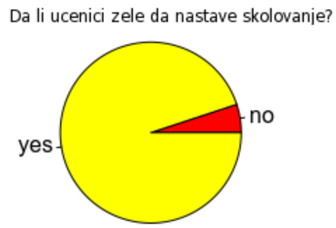
```
# Broj ucenika vs koliko sati nedeljno su ucili
ggplot(studenti,aes(x=studytime,color=sex)) + geom_histogram(
    binwidth = 0.5,fill="white") + labs(title = "Koliko ucenici
    nedeljno uce ovaj predmet")
```



Slika 2: Histogram obeležja Studytime

Analiziraćemo još jedno značajno obeležje– higher – i izračunati populacijsku srednju vrednost za one učenike koji ne žele da nastave dalje školovanje kao i za one koji žele dalje da nastave školovanje i uporediti rezultate. Sa slike 3 zaključujemo da veći broj učenika želi da nastavi školovanje.

Vidimo da upravo zbog toga histogrami se veoma razlikuju. Dolazi i do velike razlike u samoj populacijskoj srednoj vrednosti kada uporedimo ova dva klastera – kada gledamo one učenike koji žele da nastave školovanje populacijska srednja vrednost iznosi $m_{G3}^{T} = 10.608$, dok ova vrednost za one koji ne žele da nastave školovanje iznosi $m_{G3}^{F} = 6.8$

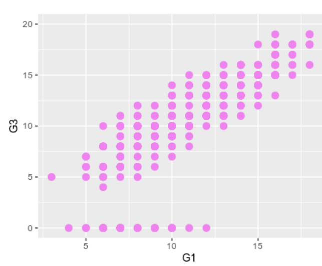


0.4

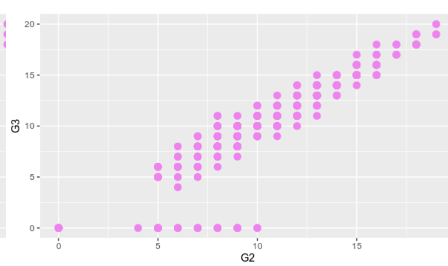
Slika 3: Pitasti dijagram odnosa broja učenika koji žele da nastave školovanje i onih koji ne žele

U nastavku je dat kod za pravljenje ovih vizuelizacija kao i za računanje populacijskih vrednosti.

```
dalje_skolovanje<-subset(studenti,higher=='yes')
ggplot(dalje_skolovanje,aes(x=G3)) + geom_histogram(binwidth = 0.5,fill="white",col="red")
mean(dalje_skolovanje$G3)
dalje_skolovanje_ne<-subset(studenti,higher=='no')
ggplot(dalje_skolovanje_ne,aes(x=G3)) + geom_histogram(binwidth = 0.5,fill="white",col="red")
mean(dalje_skolovanje_ne$G3)
```



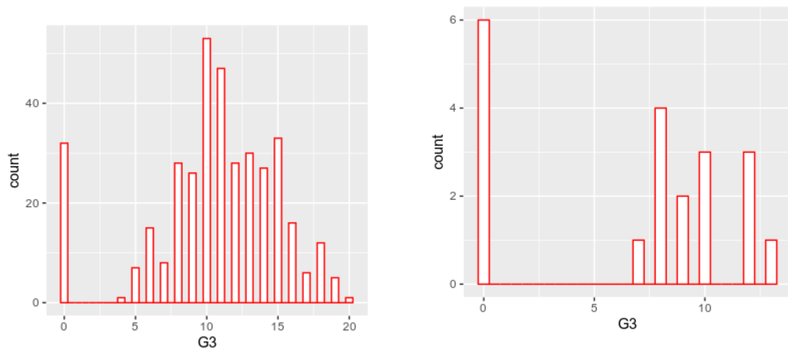
(a) Grafik zavisnosti G3 od G1



(b) Grafik zavinosti G3 od G2

Slika 4: Vizuelizacija korelacija izmedju G1 i G3 kao i G2 i G3

Poslednja stvar koju ćemo analizirati je korelacija između ocene nakon prvog semestra i finalne ocene kao i između ocene nakon drugog semestra



(a) Histogram broja poena za one koji žele da nastave školovanje

(b) Histogram broja poena za one koji ne žele da nastave školovanje

Slika 5: Vizuelizacija obeležja Higher

i finalne ocene. Sa slike se jasno vidi da neka korelacija postoji. Računamo koeficijente korelacije za oba para vrednosti.

$$corr(G1, G3) = 0.8014679$$

$$corr(G2, G3) = 0.904868$$

Zaključujemo da je finalna ocena visoko korelisana sa ocenom na kraju drugog semestra.

```
# Zavisnost g3 od g1 i g2
ggplot(studenti,aes(x=G1,y=G3)) + geom_point(col="violet",size=3)
korelacija_izmedju_g1_i_g3<-cor(studenti$G1, studenti$G3, method
    = c("pearson", "kendall", "spearman"))
korelacija_izmedju_g1_i_g3
ggplot(studenti,aes(x=G2,y=G3)) + geom_point(col="violet",size=3)
korelacija_izmedju_g2_i_g3 <-cor(studenti$G2, studenti$G3, method
    = c("pearson", "kendall", "spearman"))
korelacija_izmedju_g2_i_g3
```

3 Prost slučajan uzorak

Prost slučajan uzorak je najjednostavnija forma uzorčenja. Za njega važi da svaki od $\binom{N}{n}$, gde su N obim populacije i n obim uzorka, kod uzorka bez ponavljanja ili N^n kod uzorka sa ponavljanjem mogućih uzoraka ima podjednaku verovatnoću da bude odabran. Ovo znači da i sve jedinice populacije imaju jednaku verovatnoću da budu izvučene u uzorku, odnosno **jedinica posmatranja = jedinica uzorkovanja** . Postoje dva načina kako možemo da uzmemo prost slučajan uzorak:

- **Prost slučajan uzorak sa ponavljanjem**(*eng. SRSWR*): Možemo ga posmatrati kao da izvlačimo n nezavisnih uzoraka obima 1. Svaka jedinica se izvlači sa verovatnoćom $1/N$. U uzorku može biti dupliranih vrednosti iz populacije.

- **Prost slučajan uzorak bez ponavljanja**(*eng. SRSWOR*): Pošto nam uzorak sa ponavljanjem ne obezbeđuje dodatne informacije, obično se koristi ovaj uzorak. Verovatnoća da jedinica bude odabrana u uzorku zove se *verovatnoca prvog reda* i za SRSWOR iznosi

$$\pi_k = n/N$$

[1] Pre samog postupka uzorkovanja trebalo bi da odaberemo pogodan obim uzorka n. Da bismo odredili optimalan obim potrebno je zadati dve vrednosti, vrednost Δ koja predstavlja apsolutnu(dozvoljenu gresku) i vrednost $1-\alpha$ koja predstavlja nivo poverenja. Koristimo sledeću formulu

$$P\{|\hat{\theta} - \theta| > \Delta\} < 1 - \alpha \tag{1}$$

Odakle se dobija formula

$$n_0 = (\frac{\sigma Z_{1-\frac{\alpha}{2}}}{\Delta})^2 \tag{2}$$

Pošto nemamo σ iz prethodnih istraživanja mi ćemo koristiti tablice [3] za procenu obima uzorka. $n = 187$ Neka je S dobijen prost slučajan uzorak bez ponavljanja obima n iz populacije obima N i neka je Y obeležje od interesa čiju srednju vrednost hoćemo da ocenimo. Ocena populacijske srednje vrednosti tj uzoračka srednja vrednost data je formulom

$$\hat{m_y} = \sum_{k \in S} y_k \tag{3}$$

Ova ocena je nepristrasna, odnosno za nju vazi

$$E\hat{m}_y = m_y \tag{4}$$

Disperzija ocene populacijske srednje vrednosti je data formulom

$$D\hat{m}_y = \frac{\sigma^2 \cdot (N-n)}{N \cdot n} \tag{5}$$

Međutim pošto je σ^2 nepoznata populacijska disperzija, obicno se koristi tačkasta ocena $D\hat{m}_y$ data formulom

$$D\hat{\ }\hat{m}_y = \frac{S_n^2 \cdot (N-n)}{N \cdot n} \tag{6}$$

,gde je S_n^2 poznata uzoračka disperzija.

Pored tačkaste ocene nama su značajne još i intervalne ocene. Mi određujemo interval I za koga važi

$$P\{\theta \in I\} = 1 - \alpha \tag{7}$$

Aproksimativni interval poverenja za vrednost m_y je

$$[\hat{m}_y - Z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{S_n^2}{n} \cdot (1 - \frac{n}{N}}, \hat{m}_y + Z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{S_n^2}{n} \cdot (1 - \frac{n}{N}}] \tag{8}$$

Implementacija ovih formula za našu bazu i naše obeležje od interesa G3 je data u nastavku

```
n=187 # odredjeno na osnovu tablice
N=length(studenti$sex)
N
indeks<-sample(N,n,replace=FALSE)
uzorak = studenti$G3[indeks]
length(uzorak)
# Ocenjujemo populacijsku srednju vrednost finalne ocene
xn_ocena <- mean(uzorak)
xn_ocena
# Ocenjujemo disprziju uzoracke sredine
Sn2<-var(uzorak)
D_xn_ocena <- (N-n)*Sn2/(N*n)
D_xn_ocena
# Hocemo da nadjemo 95% aproksimativni interval poverenja
# Koristimo umesto populacijske srednje vrednosti uzoracku
alpha<-0.05
z<-qnorm(1-alpha/2)
intervalPoverenja<-c(xn_ocena - z*sqrt(D_xn_ocena),xn_ocena - z*
    sqrt(D_xn_ocena))
intervalPoverenja
```

Vrednost uzoračke sredine je $\hat{m}_y = 10.70053$, dok je ocena dizperzija $D\hat{\ } m_y = 0.05240131$, a interval poverenja $I_{m_y} = [10.25187, 10.25187]$ što je i očekivana vrednost ako uzmemo u obzir da raspodela ocena teži da ima normalnu raspodelu kao i srednju vrednost koja je kao medijana.

3.1 Količničko ocenjivanje

Kod količnickog ocenjivanja potrebno je pronaći obeležje koje ima približno linearnu vezu sa našim obeležjem od interesa i čiju vrednosti možemo odrediti na proizvoljnoj jedinici u populaciji. Takođe, vrednost totala ovog pomoćnog obeležja X, τ_x , mora biti poznata U delu 1.1 smo videli da postoji korelacija između obeležja $G2$ i $G3$ koju ćemo iskoristiti za količničko ocenjivanje.

Definišemo populacijski količnik

$$B = \frac{m_y}{m_x} \tag{9}$$

,gde je Y obeležje od interesa, a X pomoćno obeležje. Želimo da ocenimo parametar B kao i parametar m_y. Za ocenu populacijskog količnika koristimo uzorački količnik

$$b = \frac{\hat{m}_y}{\hat{m}_x} \tag{10}$$

, a za ocenu populacijske srednje vrednosti koristimo

$$\hat{m}_y{}^r = b \cdot m_x \tag{11}$$

Ova ocena nije nepristrasna, njena pristrasnost iznosi $-cov(b, \hat{m}_x)$, ali je zanemarljivo mala ako su visoko korelisani ili ako je n veliko. Ocena disperzije je data formulom

$$\hat{D}_{\hat{m_y}^r} = \frac{S_e^2}{n} \cdot (1 - \frac{n}{N}) \tag{12}$$

gde je

$$S_e^2 = \frac{1}{n-1} \sum_{k \in S} (y_k - b \cdot x_k)^2 \tag{13}$$

U kodu u nastavku su implementirane ove formule za prost slučajan uzorak S koji je gore naveden.

```
# Populacijska srednja vrednost pomocnog obelezja
mx<-mean(studenti$G2)
# Vadimo vrednosti obelezja G2 za uzorak
g2_iz_uzorka <- studenti$G2[indeks]
# Uzoracka srednja vrednost za pomocno obelezje
mx_ocenjeno <- mean(g2_iz_uzorka)
mx_ocenjeno
# Preimenovali smo samo uzoracku srednju vrednost obelezja od
    interesa
my_ocenjeno <- xn_ocena
# Racunamo uzoracki kolicnik
b_ocena <- mx_ocenjeno / my_ocenjeno
# Kolicnicka ocena za obelezje G3
xn_kol_ocena <- b_ocena * mx
xn_kol_ocena
# Ocena disperzije kolicnicke ocene
Se_2 <- sum((uzorak-b_ocena*g2_iz_uzorka)^2) / (n-1)
D_xn_kol <- (Se_2 / n)* (1-n/N)
D_xn_kol
```

Vrednost količničke ocene populacijske sredine je $\hat{m}_y^r = 11.07232$, dok je ocena dizperzija $\hat{D}_{\hat{m}_y^r} = 0.01163492$.

4 Stratifikovani uzorak

Straficikacija je podela populacije na potpopulacije n – stratume – na osnovu nekih dodatnih informacije. Kriterijum za podelu nam je obično neko obeležje koje je u vezi sa obeležjem od interesa. Stratumi su međusobno disjunktni tj moraju pokriti celu populaciju. Nakon same stratifikacije iz svakog stratuma se bira određeni broj jedinica. Nije neophodno da se isti metod odabira uzorka primeni na svaki stratum. Ovakav uzorak se naziva stratifikovani uzorak. Uzorak koji ćemo mi ovde primeniti je stratifikovani slučajan uzorak – iz svakog stratuma ćemo odabrati prost slučajan uzorak bez ponavljanja. Prilikom podele na stratume treba voditi računa da bude ispunjena relativna *homogenost* unutar stratuma kao i relativna *heterogenost* između stratuma. Napravićemo dve podele po stratumima i za svaku od podela ćemo izračunati ocenjenu populacijsku vrednost za obeležje $G3$ i to na osnovu:

- pola
- da li žele da nastave dalje školovanje

Drugo pitanje koje nam se nameće je pitanje rasporeda uzorka po stratumima. Koristićemo proporcionalni raspored za koji važi da je broj jedinica koje se biraju za uzorak proprocionalan broju jedinica u stratumu odnosno za h-ti stratum važi:

$$n_h = \frac{n \cdot N_h}{N} \tag{14}$$

gde je N_h obim h-tog stratuma. Nepristrasna ocena za populacijsku srednju vrednost je data:

$$\hat{m_y}^{str} = \frac{1}{N} \cdot \sum_{h=1}^{L} N_h \cdot \hat{m}_{y_h} \tag{15}$$

gde su L – broj stratuma, $\hat{m}_{y_h}$ – uzoračka sredina h-tog stratuma. Ocena disperzije je jednaka:

$$\hat{D}(\hat{m_y}^{str}) = \frac{1}{N^2} \cdot \sum_{h=1}^{L} \frac{N_h^2 \cdot S_{n_h}^2}{n_h} \cdot (1 - \frac{n}{N}) \tag{16}$$

Kod za implementaciju ovih formula i formiranje stratuma za ova dva obeležja je dat u nastavku

```
# pol
stratumF <- subset(studenti,sex=='F')
N_f = length(stratumF$sex)
stratumM <- subset(studenti,sex=='M')
N_m = length(stratumM$sex)
n_momci <- (n/N)*N_m
n_devojke <- (n/N)*N_f
n_momci+n_devojke == n
# uzorkujemo stratumF
indeksF <-sample(N_f,n_devojke,replace = FALSE)
uzorakF<- stratumF$G3[indeksF]
xn_devojke<-mean(uzorakF)
SnF_2<-var(uzorakF)
D_xn_devojke <- (N_f-n_devojke)*SnF_2/(N_f*n_devojke)
# uzorukujemo stratumM
indeksM <-sample(N_m,n_momci,replace = FALSE)
uzorakM<- stratumM$G3[indeksM]
xn_momci<-mean(uzorakM)
SnM_2<-var(uzorakM)
```

```
D_xn_momci <- (N_m-n_momci)*SnM_2/(N_m*n_momci)
# ocenjujemo populacijsku sredinu
xn_strat <-(xn_devojke*N_f + xn_momci*N_m)/N
xn_strat
# ocenjujemo disperziju
disperzije <-c(D_xn_momci,D_xn_devojke)
Nh<-c(N_m,N_f)
nh<-c(n_momci,n_devojke)
Sn_2<-c(SnM_2,SnF_2)
D_xn_strat<-(1/N^2)*sum(((Nh^2*Sn_2)/nh)*(1-n/N))
D_xn_strat

# da li zele da nastave skolovanje
stratumYes <- subset(studenti,higher=='yes')
N_yes = length(stratumYes$higher)
stratumNo<- subset(studenti,higher=='no')
N_no = length(stratumNo$higher)
N_yes+N_no == N
n_yes <- round((n/N)*N_yes)
n_no <- round((n/N)*N_no)
n_yes+n_no == n
# uzorkujemo stratumYes
indeksYes <-sample(N_yes,n_yes,replace = FALSE)
uzorakYes<- stratumYes$G3[indeksYes]
xn_yes<-mean(uzorakYes)
SnYes_2<-var(uzorakYes)
D_xn_Yes <- (N_yes-n_yes)*SnYes_2/(N_yes*n_yes)
# uzorukujemo stratumNo
indeksNo <-sample(N_no,n_no,replace = FALSE)
uzorakNo<- stratumM$G3[indeksNo]
xn_no<-mean(uzorakNo)
SnNo_2<-var(uzorakNo)
D_xn_No <- (N_no-n_no)*SnNo_2/(N_no*n_no)
# ocenjujemo populacijsku sredinu
xn_strat <-(xn_yes*N_yes + xn_no*N_no)/N
xn_strat
# ocenjujemo disperziju
disperzije <-c(D_xn_Yes,D_xn_No)
Nh<-c(N_yes,N_no)
nh<-c(n_yes,n_no)
Sn_2<-c(SnYes_2,SnNo_2)
D_xn_strat<-(1/N^2)*sum(((Nh^2*Sn_2)/nh)*(1-n/N))
D_xn_strat
```

Vrednosti koje dobijamo prilikom ovog izvršavanja su date u tabeli ispod.

obeležje	$\hat{x_n}^{str}$	$\hat{D}(\hat{m_y}^{str})$
pol	10.10241	0.06542043
dalje školovanje	10.89326	0.06005498

Kao što se vidi iz tabele obe podele po stratumima su dali dobre rezultate procene srednje vrednosti. Disperzija je manja prilikom korišćenja obeležja higher(dalje školovanje). Obe ocene su nepristrasne pa za poređenje bolje ocene se koristi disperzija. Zaključujemo da bolju ocenu daje stratifikacija po obeležju *higher*.

5 Grupni uzorak

Kod grupnog uzorka važi da se razlikuju jedinice posmatranja od jedinica uzorkovanja. Jedinice uzorkovanja su nam *primarne jedinice* odnosno *grupe*, dok su *sekundarne jedinice*, tj. jedinice posmatranja, zapravo entiteti unutar tih grupa.
Na osnovu nekog kriterijuma vrši se podela na grupe, a zatim se nekim metodom uzorkovanja vrši odabir grupa. Iz grupe posmatramo sve entitete. Ono na šta treba voditi računa prilikom odabira grupa je da one budu što sličnije međusobno odnosno da ih odlikuje *relativna homogenost*, dok unutar grupe entiteti treba da budu što različitiji odnosno da ih odlikuje *relativna* heterogenost.
Prvi kriterijum za podelu na grupe koji se kod nas nameće je škola koju pohađaju. Naime, pretpostavićemo da su svake škole relativno slične po ocenama i uspehu, dok su u školama entiteti odnosno učenici heterogeni.Pošto je velika razlika u obimu grupa, za školu GP obim je $M_{GP} = 349$, dok je obim grupe MS $M_{MS} = 46$ prirodno se nameće uzorčenje sa nejednakim verovatnoćama proporcionalnim veličini grupe.
Neka nam je N broj primarnih jedinica, a n obim uzorka primarnih jedinica tj. grupa. Takođe, neka su M_l broj sekundarnih jedinica u h-toj grupi i $M = \sum_{l=1}^{N} M_l$ obim populacije. Verovatnoća odabira l-te grupe data je formulom:

$$\psi_l = \frac{M_l}{M} \tag{17}$$

Nepristrasna ocena $\hat{m}_y^{grp}$ data je formulom:

$$\hat{m}_y^{grp} = \frac{1}{n} \cdot \sum_{i \in S} \frac{t_l}{M_l} \tag{18}$$

gde su redom S – neuređen skup kardinalosti n, t_i – total primarne jedinice i. Ocena disperzije data je formulom:

$$\hat{D}(\hat{m}_y^{grp}) = \frac{1}{n} \cdot \frac{1}{n-1} \sum_{i \in S} (m_i - \frac{\hat{t}_y^{grp}}{M})^2 \tag{19}$$

U narednom kodu je data implementacija ovih formula.

```
skola = c('GP','MS')
N = 2
Mu = 395
n = 1
skola1 <- subset(studenti,school == 'GP')
Mgp <- length(skola1$G3)
Mgp
skola2 <- subset(studenti,school == 'MS')
Mms <- length(skola2$G3)
Mms
M_skole <-c(Mgp,Mms)
odabranaSkola = sample(skola,1,prob = c(Mgp/M,Mms/M))
odabranaSkola
uzorak = subset(studenti,school == odabranaSkola)
Ml = length(uzorak$G3)
# Ocenjujemo populacijsku srednju vrednost
t_n_grp = Mu/n*(sum(uzorak$G3)/Ml)
x_n_grp = t_n_grp / Mu
x_n_grp
```

Vrednosti koje dobijamo za ocenu populacijske vrednosti date su u narednoj tabeli.

grupa	$\hat{m_y}^{grp}$
GP	10.48997
MS	9.847826

Naime, velika veličina grupe dovodi do smanjenje uštede, a ta ušteda je glavna prednost ovog uzorka. Samim tim uzorčenje ovakvih grupa nije efikasan metod. U samoj bazi su date samo dve škole koje od kojih smo birali samo jednu.

Dalje, drugi argument koji koristimo za podelu na grupe je argument *Age*. Uzećemo 4 grupe kako bismo demonstritali. Ovog puta uzorak koji koristimo je prost slučajan uzorak primarnih jedinica odnosno grupa. Formula za ocenu populacijske srednje vrednosti kao i ocena disperzije je data u nastavku. Kao i za sve ostale, tako i za ovu važi nepristrasnost ocene.

$$\hat{m}_y^{grp} = \frac{N}{n \cdot M} \sum_{l \in S} t_l \tag{20}$$

$$\hat{D}(\hat{m}_y^{grp}) = \frac{N^2}{n} \cdot (1 - \frac{n}{N}) \cdot \frac{1}{M^2} \cdot S_t^2 \tag{21}$$

, gde su

$$S_t^2 = \frac{1}{n} \sum_{i \in S} (t_l - \hat{y}_t)^2 \tag{22}$$

$$\hat{y}_t = \frac{1}{n} \sum_{i \in S} t_i \tag{23}$$

Kod za implementaciju ovih vrednosti dat je u nastavku.

```

# Grupni uzorak -- razredi

n_razredi = 4
N_razreda = max(studenti$age) - min(studenti$age)
N_razreda
odabraniRazredi = sample(15:22,n_razredi,replace=FALSE)
odabraniRazredi
grupa1 = subset(studenti,age==odabraniRazredi[1])
grupa2 = subset(studenti,age==odabraniRazredi[2])
grupa3 = subset(studenti,age==odabraniRazredi[3])
grupa4 = subset(studenti,age==odabraniRazredi[4])
# Ovo nam je prost slucajan uzorak
t_xn_grupa = (N_razreda/n_razredi)*(sum(grupa1$G3) + sum(grupa2$G3) + sum(grupa3$G3) + sum(grupa4$G3))
x_xn_grupa = t_xn_grupa / Mu
x_xn_grupa
# Racunamo disperziju
X_tau = (sum(grupa1$G3) + sum(grupa2$G3) + sum(grupa3$G3) + sum(grupa4$G3)) / n_razredi
X_tau
```

```
s2_tau = ((sum(grupa1$G3) - X_tau)^2 + (sum(grupa2$G3) - X_tau)^2
    + (sum(grupa3$G3) - X_tau)^2+(sum(grupa4$G3) - X_tau)^2)/(n_
    razredi -1)
D_xn_grupa = (1/Mu^2)*((N_razreda^2)/n_razredi)*(1-n_razredi/N_
    razreda)*s2_tau
D_xn_grupa
```

Vrednost koju smo dobili je $\hat{m}_y^{grp} = 8.989241$ što je lošija ocena populacijske srednje vrednosti.

6 Zaključak

Kao što smo videli u prethodnim poglavljima svaki od uzoraka je lepo procenio populacijsku srednju vrednost našeg obeležja od interesa. Prost slučajan uzorak je dao najbolju ocenu obeležja od interesa.

Literatura

[1] Sharon L. Lohr. *Sampling Design And Analysis*. Duxbury Press 1999. 23-24 str.

[2] Lenka Glavaš. *Slajdovi sa predavanja za kurs Uvod u Teoriju Uzorka*

[3] Glenn D. Israel. *Determining Sample Size*. University Of Florida 1992.