UNIVERZITET U NOVOM SADU

PRIRODNO – MATEMATIČKI FAKULTET

DEPARTMAN ZA MATEMATIKU I INFORMATIKU

Dunja Arsić

KLASTER UZORCI SA JEDNAKIM VEROVATNOĆAMA

– master rad –

Mentor: Prof. dr Sanja Rapajić

Novi Sad, 2012.

Zahvaljujem se svim profesorima i asistentima Prirodno-matematičkog fakulteta u Novom Sadu sa kojima sam sarađivala tokom studija na znanjima koja su mi nesebično pružena.

Takođe, zahvalila bih se članovima komisije, prof. dr Zorani Lužanin i doc. dr Sanji Konjik na saradnji, kao i prof. dr Nataši Krejić na sugestijama i dragocenoj podršci.

Veliku zahvalnost dugujem izuzetnom profesoru i još izuzetnijoj osobi, svom mentoru, prof. dr Sanji Rapajić na savetima, pomoći i podršci pruženoj kako prilikom izrade ovog rada tako i tokom čitavog studiranja.

Posebno, najveću zahvalnost dugujem svojoj porodici, onima koji me svakodnevno okružuju i čine osobom kakva jesam, ljudima koje najviše volim...

Dunja Arsić

Sadržaj

1 Uvod **3**
- 1.1 Uzorkovanje . . . 5
  - 1.1.1 Osnove verovatnosnog uzorkovanja . . . 8

2 Prost slučajan uzorak **13**

3 Ocena količnika **18**
- 3.1 Razlozi za korišćenje ocene količnika . . . 19
- 3.2 Pristrasnost i srednje kvadratna greška ocenjivača količnika . . . 23
  - 3.2.1 Prednosti ocene količnika . . . 27

4 Klaster uzorci sa jednakim verovatnoćama **29**
- 4.1 Oznake u klaster uzorkovanju . . . 33
- 4.2 Jednofazno klaster uzorkovanje . . . 34
  - 4.2.1 Klasteri jednakih veličina - ocenjivanje . . . 34
  - 4.2.2 Klasteri jednakih veličina - teorija . . . 37
  - 4.2.3 Klasteri različitih veličina . . . 42
- 4.3 Dvofazno klaster uzorkovanje . . . 45
- 4.4 Korišćenje težina u klaster uzorcima . . . 51
- 4.5 Dizajniranje klaster uzorka . . . 52
- 4.6 Sistematsko uzorkovanje . . . 56

5 Primena klaster uzorka **62**
- 5.1 Jednofazni klaster uzorak - klasteri su departmani . . . 63
- 5.2 Jednofazni klaster uzorak - klasteri su godine studija . . . 67
- 5.3 Dvofazni klaster uzorak - klasteri su departmani . . . 69
- 5.4 Dvofazni klaster uzorak - klasteri su godine studija . . . 73

6 Zaključak **76**

1

Uvod

Populacija je celokupna kolekcija objekata u kojoj se može vršiti ispitivanje neke karakteristike tj. nekog obeležja. Ukoliko je populacija malobrojna, ona se može izučavati u celini. Međutim, ako ona sadrži veliki broj elemenata, ispitivanje cele populacije je skupo, dugotrajno, u nekim slučajevima može biti destruktivno, a ponekad je čak i principijelno nemoguće. Iz tog razloga se obično bira podskup populacije koji se naziva uzorak, na kome se vrši ispitivanje. Ideja je da pokušamo da izvedemo zaključak o celoj populaciji na osnovu analize izabranog uzorka.

Oblast statistike koja se bavi proučavanjem izbora uzorka i ocenjivanjem odgovarajućih parametara populacije naziva se teorija uzoraka. Postoje mnogobrojne tehnike i načini odabira uzorka, a samim tim i razne vrste uzoraka. U ovom radu posmatran je klaster uzorak, koji uz prost slučajan uzorak i stratifikovani uzorak spada u osnovne tipove verovatnosnih uzoraka. Predstavljeni su i analizirani ocenjivači ukupne i srednje vrednosti populacije kod klaster uzorka, sa akcentom na nepristrasnim ocenama i ocenama količnika. Osim toga, prikazan je i sistematski uzorak kao specijalan slučaj klaster uzorka.

Klaster uzorak se koristi kada su članovi populacije grupisani u prirodne klastere ili kada je nepoznat spisak svih članova populacije, tj. kada je formiranje uzoračkog okvira komplikovano, skupo ili nemoguće. Klaster uzorak čine slučajno izabrane primarne uzoračke jedinice (klasteri), u okviru kojih se biraju i ispituju svi ili samo neki elementi koji predstavljaju sekundarne uzoračke jedinice. U zavisnosti od toga, postoji jednofazni i dvofazni klaster uzorak.

Jednofazni klaster uzorak se koristi kad su troškovi uzorkovanja sekundarnih uzoračkih jedinica zanemarljivi u poređenju sa troškovima uzorkovanja klastera. U jednofaznom klaster uzorku ispituju se svi elementi unutar

izabranih klastera. Pošto oni mogu biti veoma slični, ispitivanje svih tih elemenata može biti skupo i nepotrebno. U takvim situacijama jeftinije je biranje poduzorka iz svakog izabranog klastera i tada je reč o dvofaznom klaster uzorku.

Ocene populacionih veličina dobijene iz klaster uzorka su u opštem slučaju manje precizne od ocena dobijenih iz prostog slučajnog uzorka, jer su obično elementi istog klastera sličniji nego slučajno odabrani elementi iz cele populacije. Uzorkovanjem sličnih elemenata iz jednog klastera ne dobijaju se nove informacije, što smanjuje preciznost ocena populacionih parametara. Međutim, i pored toga, klaster uzorci se veoma često koriste u praksi kada su u pitanju velika istraživanja, jer je obično mnogo jeftinije i jednostavnije uzorkovati elemente unutar klastera nego birati prost slučajan uzorak iz cele populacije.

U prvom, uvodnom delu rada prikazane su osnove verovatnosnog uzorkovanja. U drugom poglavlju predstavljen je prost slučajan uzorak, kao osnovni oblik verovatnosnog uzorkovanja koji daje teorijsku osnovu za složenije oblike. Ocena količnika je opisana u trećem poglavlju. Četvrto poglavlje predstavlja ključni deo rada u kojem je detaljno obradjeno klaster uzorkovanje. U petom poglavlju prikazana je primena klaster uzorka na konkretnom realnom problemu.

1.1 Uzorkovanje

Ispitivanje čitave populacije je, u opštem slučaju, nemoguće kada je reč o brojnim populacijama. Usled nedostatka vremena i novca ili usled brojnosti populacije, najčešće nismo u stanju da ispitamo svakog stanovnika jedne države. Sa druge strane, istraživanje svih geografskih područja na Zemlji je fizički nemoguće. Iz tog razloga, radi dobijanja željenih informacija o čitavoj populaciji, istraživači se oslanjaju na uzorkovanje. Uzorkovanje predstavlja proces selekcije podskupa jedinica iz cele populacije radi ocenjivanja karakteristika čitave populacije.

Osnovni pojmovi iz oblasti uzorkovanja su:

Jedinica posmatranja (element) – objekat čije karakteristike se ispituju. Na primer, u proučavanju ljudske populacije jedinice posmatranja su pojedinci, tj. osobe.

Ciljna populacija – svaka kolekcija ljudi, biljaka, životinja ili stvari u kojoj se vrši ispitivanje neke karakteristike. Njeno definisanje predstavlja važan deo istraživanja pošto odabir ciljne populacije ima veliki uticaj na kasnije dobijene statističke rezultate.

Uzorak – podskup populacije. Analizom izabranog uzorka izvodimo zaključke o celoj populaciji.

Uzoračka populacija – skup svih mogućih jedinica posmatranja koje imaju šansu da se nađu u uzorku. Dakle, to je populacija iz koje se uzima uzorak.

Uzoračka jedinica – jedinica koja se uzorkuje. Na primer, u istraživanju određivanja prosečnog prihoda po osobi, domaćinstva mogu biti uzoračke jedinice, a pojedinci su u tom slučaju jedinice posmatranja.

Uzorački okvir – spisak svih uzoračkih jedinica. Na primer, u telefonskom istraživanju Novog Sada, uzorački okvir je lista svih telefonskih brojeva u Novom Sadu.

Ključ svakog uspešnog istraživanja je dobra priprema. Prema tome, ukoliko se uzorkovanje dovoljno dobro pripremi i sprovede, ono može u velikoj meri uštedeti vreme, novac i trud, pružajući tačne, precizne, pouzdane i korisne rezultate. Takođe, pored smanjenja troškova, brzina prikupljanja podataka je jedna od najvećih prednosti uzorkovanja.

Prilikom svakog istraživanja javljaju se greške koje se dele na uzoračke i neuzoračke. Razumevanje razlika između ovih grešaka je od velike važnosti pri opredeljivanju za popis ili za uzorak, kao i pri odabiru odgovarajućih pro-

cedura u slučaju da se odlučimo za uzorkovanje. Da bi rezultati istraživanja bili što pouzdaniji, potrebno je minimizirati sve vrste grešaka.

Uzoračka greška je greška koja nastaje kao posledica korišćenja uzorka umesto ispitivanja cele populacije. Ona predstavlja razliku između ocene nekog obeležja dobijene na osnovu uzorka i prave vrednosti obeležja dobijene na celoj populaciji. Uzoračka greška se razlikuje od uzorka do uzorka.

U većini istraživanja uzoračke greške su zanemarljive u odnosu na neuzoračke. Uzoračka greška odražava preciznost ocene, a neuzoračka greška validnost ocene. Neuzoračke greške su greške koje se ne pripisuju variranju od uzorka do uzorka. To su greške koje potiču od načina uzorkovanja kojim se dobijaju ocene koje se sistematski razlikuju od pravih vrednosti obeležja populacije. Neuzoračke greške su razne vrste pristrasnosti koje nastaju u procesu istraživanja, kao na primer pristrasnost prilikom odabira i pristrasnost prilikom merenja.

U današnje vreme postoji veliki broj loše odrađenih istraživanja, pa su mnogi ljudi prilično skeptični prema većini istraživanja. Neki smatraju da je uzorkovanje loše i da istraživanje treba vršiti isključivo na celoj populaciji, tj. treba raditi kompletan popis. Za male populacije to nije problem i kompletnim popisom se eliminiše uzoračka greška. Međutim, popis ne može isključiti neuzoračke greške. Najveći uzroci grešaka u istraživanjima su nepokrivenost, neodziv i nemarnost pri sakupljanju podataka. U opštem slučaju, kompletan popis populacije iziskuje previše vremena i novca, ponekad može biti destruktivan, a ne eliminiše sve greške. Dakle, čak i kad smo u prilici da ispitujemo celu populaciju, često se odlučujemo da selektujemo samo njen manji deo (uzorak) i da na osnovu njega donosimo odluke vezane za celu populaciju. Postoje mnogobrojni razlozi za to. Uzorkovanje može obezbediti pouzdane informacije za mnogo manje novca od popisa. Podaci dobijeni na osnovu uzorka se brže prikupljaju, pa se i ocene veličina objavljuju blagovremeno. Osim toga, ocene bazirane na uzorku su često tačnije od onih dobijenih na osnovu popisa, jer se veća pažnja posvećuje kvalitetu podataka i obučavanju osoblja koje sprovodi istraživanje, što bitno smanjuje greške prilikom istraživanja. Mnogo je bolje imati dobra merenja na reprezentativnom uzorku, nego pristrasna ili nepouzdana merenja na celoj populaciji.

Naša saznanja, stavovi i dela su u velikoj meri zasnovani na uzorcima, što možemo videti kako u naučnim istraživanjima, tako i u svakodnevnom životu. Iako se veruje da je uzorkovanje oduvek bilo deo ljudske istorije, mnoge uzoračke procedure koje se danas koriste imaju relativno kratku istoriju. Neuspeh velikih kompanija u istraživanju javnog mnjenja povodom predviđanja pobednika na američkim predsedničkim izborima 1948. godine, motivisao ih je da u svoje procedure predviđanja prvi put uključe verovat-

nosno uzorkovanje, koje je ubrzo postalo dominantan alat za ocenjivanje parametara populacije.

Verovatnosno uzorkovanje je vrsta uzorkovanja kod kojeg svaki element ciljne populacije ima poznatu, određenu verovatnoću da bude izabran u uzorak. Ove verovatnoće odabira elemenata mogu biti jednake ili različite. U slučaju da svi elementi populacije imaju jednake verovatnoće odabira, reč je o prostim verovatnosnim uzorcima.
Nasuprot toga, neverovatnosno uzorkovanje je vrsta uzorkovanja kod kojeg neki elementi populacije nemaju šansu da budu izabrani ili verovatnoće izbora elemenata nisu poznate.

Tri osnovna tipa verovatnosnog uzorka su: prost slučajan, stratifikovan i klaster uzorak.

Prost slučajan uzorak

Prost slučajan uzorak (eng. Simple Random Sample - SRS) je najjednostavniji oblik verovatnosnog uzorka. Njegova osnovna karakteristika je da svi članovi populacije imaju jednaku šansu da budu uključeni u uzorak, a takođe i svaka moguća kombinacija datog broja članova populacije sa istom verovatnoćom može biti uključena u uzorak. Ovaj tip uzorka predstavlja osnovu za složenije tipove verovatnosnih uzoraka.

Stratifikovani uzorak

Stratifikovani uzorak se koristi kad je populacija podeljena na podgrupe koje se nazivaju stratumi. Iz svakog stratuma se nezavisno bira prost slučajan uzorak. Stratumi su najčešće podgrupe elemenata sličnih ili istih osobina. Stratumi, na primer, mogu biti starosne grupe ljudi, različite vrste terena ili različite veličine firmi. Elementi unutar istog stratuma su obično sličniji nego proizvoljno izabrani elementi iz cele populacije. Pošto se u uzorku nalaze elementi iz svakog stratuma, stratifikacija u opštem slučaju povećava preciznost.

Klaster uzorak

Klaster uzorak je uzorak u kojem su posmatrane jedinice populacije grupisane u veće uzoračke jedinice koje se nazivaju klasteri. Bira se prost slučajan uzorak klastera i ispituju se svi ili samo neki elementi izabranih klastera.

Na primer, umesto anketiranja svih sportista Novog Sada, čiji spisak nemamo, uzima se prost slučajan uzorak sportskih klubova, a zatim se anke-

tiraju svi ili neki od sportista u izabranim sportskim klubovima. U ovom primeru, sportski klubovi su klasteri, a sportisti jedinice posmatranja.

Primer 1.1 *Pretpostavimo da želimo da ocenimo prosečno vreme koje profesori univerziteta utroše na pregledanje pismenih radova tokom junskog ispitnog roka.*

Jedan način je da koristimo prost slučajan uzorak. U tom slučaju je potrebno da posedujemo spisak svih profesora univerziteta i tada slučajno biramo profesore sa tog spiska. Drugi način je korišćenje stratifikovanog uzorka. Tada nam je potreban spisak svih fakulteta i profesora koji rade na njima, pa u okviru svakog fakulteta biramo prost slučajan uzorak. Treći način je slučajno biranje nekih fakulteta i ispitivanje svih profesora sa izabranih fakulteta i to je klaster uzorak. Uočava se da je u sva tri slučaja prisutan slučajan izbor jedinica u uzorak. Kod prostog slučajnog uzorka jedinice posmatranja su slučajno izabrane iz populacije. Kod stratifikovanog uzorka se jedinice posmatranja slučajno biraju iz svakog stratuma, dok se kod klaster uzorka klasteri slučajno biraju. □

1.1.1 Osnove verovatnosnog uzorkovanja

Pretpostavimo da populacija ima N elemenata (jedinica populacije), čiji su indeksi $1, 2, ..., N$, tj. indeksni skup cele populacije je $U = \{1, 2, ..., N\}$. Svakom elementu populacije je pridružena vrednost obeležja y koje posmatramo. Iz populacije se bira uzorak veličine n. Taj uzorak ćemo označavati sa $\mathcal{S}$ i on predstavlja n-točlani podskup populacije.

U verovatnosnom uzorkovanju svaki uzorak $\mathcal{S}$ iz populacije ima poznatu verovatnoću $P(\mathcal{S})$ sa kojom može biti izabran i suma verovatnoća svih mogućih uzoraka je 1. Kako je u verovatnosnom uzorkovanju za svaki mogući uzorak poznata verovatnoća sa kojom on može biti izabran, onda i svaka jedinica u populaciji ima poznatu verovatnoću sa kojom se može naći u uzorku i to je

$$P(i\text{-ta jedinica je u uzorku}) = \pi_i.$$

π_i se računa kao suma verovatnoća svih mogućih uzoraka koji sadrže jedinicu i. Verovatnoće π_i su poznate pre početka uzorkovanja i pretpostavljamo da je $\pi_i > 0$, za svaku jedinicu populacije. U praksi, ipak, nije moguće ispisati sve moguće uzorke i izračunati verovatnoće s kojima mogu biti izabrani.

Primer 1.2 *Neka je data populacija čiji je indeksni skup*

$$U = \{1, 2, 3, 4, 5, 6, 7, 8\}.$$

Neka je

y_i – *vrednost obeležja i-te jedinice.*

Broj mogućih uzoraka veličine 4 je $\binom{8}{4} = 70$. *Dakle, postoji 70 različitih uzoraka obima 4 koji se mogu dobiti iz ove populacije.*
Neka je $P(\mathcal{S}) = \frac{1}{70}$, *za svaki mogući uzorak veličine 4 iz populacije. Svaka jedinica se javlja u tačno 35 uzoraka pa je*

$$\pi_i = \sum_{i\in\mathcal{S}} P(\mathcal{S}) = \sum_{i\in\mathcal{S}} \tfrac{1}{70} = 35(\tfrac{1}{70}) = \tfrac{1}{2}, \quad i = 1, 2, ..., 8. \qquad \square$$

Odabir verovatnosnog uzorka je komplikovaniji od odabira pogodnog uzorka, ali verovatnosno uzorkovanje garantuje da se svaka jedinica populacije može pojaviti u uzorku i obezbeđuje informaciju pomoću koje se može proceniti preciznost statistike izračunate iz uzorka. Verovatnosno uzorkovanje omogućava da se koristeći uzorak relativno malog obima, mogu izvesti pouzdani zaključci o proizvoljno velikoj populaciji.

Što je uzorak manji i cena eksperimenta niža, zaključci su obično manje pouzdani i obrnuto. Zato je potrebno da budemo u mogućnosti da merimo koliko su nam zaključci i ocene dobri i koliko bi nas koštalo da ih učinimo boljim. Kod verovatnosnog uzorkovanja možemo da merimo koliko je naš uzorak dobar, tj. koliko često uzorci dosežu neki kriterijum reprezentativnosti.

Većina rezultata u uzorkovanju se oslanja na uzoračku raspodelu statistike, raspodelu različitih vrednosti statistike dobijenu u procesu uzimanja svih mogućih uzoraka iz populacije. Uzoračka raspodela je primer diskretne raspodele.

Neka posmatrana populacija ima N jedinica i neka je y_i vrednost obeležja i-te jedinice. Pretpostavimo da želimo da koristimo uzorak da bismo ocenili ukupnu vrednost populacije

$$t = \sum_{i=1}^{N} y_i.$$

Ocenjivač za t može biti

$$\hat{t}_{\mathcal{S}} = N\bar{y}_{\mathcal{S}}$$

gde je $\bar{y}_{\mathcal{S}}$ srednja vrednost y_i-ova u izabranom uzorku $\mathcal{S}$.
Uzoračka raspodela za $\hat{t}$ je

$$P\{\hat{t} = k\} = \sum_{\mathcal{S}:\hat{t}_{\mathcal{S}}=k} P(\mathcal{S}).$$

Sumiranje se vrši po svim uzorcima $\mathcal{S}$ za koje je $\hat{t}_{\mathcal{S}} = k$. Verovatnoća $P(\mathcal{S})$ sa kojom biramo uzorak $\mathcal{S}$ je poznata, pošto je reč o verovatnosnom uzorku.

Očekivana vrednost od $\hat{t}$, $E(\hat{t})$, je sredina uzoračke raspodele od $\hat{t}$

$$E(\hat{t}) = \sum_{\mathcal{S}} P(\mathcal{S})\hat{t}_{\mathcal{S}} = \sum_{k} kP\{\hat{t} = k\}.$$

Pristrasnost ocenjivača t je

$$Bias(\hat{t}) = E(\hat{t}) - t.$$

Ukoliko je $Bias(\hat{t}) = 0$, kažemo da je $\hat{t}$ nepristrasan ocenjivač za t.

Moramo napomenuti da upravo prikazana matematička definicija pristrasnosti ocenjivača nije isto što i pristrasnost prilikom odabira ili pristrasnost prilikom merenja. Sve tri vrste pristrasnosti predstavljaju odstupanje od tačne vrednosti parametra populacije, ali potiču od različitih izvora. Recimo da želimo da ocenimo prosečnu visinu košarkaša reprezentacije Srbije. Pristrasnost prilikom odabira bi se dogodila ukoliko bismo uzeli pogodan uzorak, od na primer, najviših košarkaša reprezentacije. Do pristrasnosti prilikom merenja bi došlo ukoliko bi naš metar prikazivao pogrešne mere, na primer, ukoliko bi prikazivao $2cm$ više od stvarne visine. Pristrasnost ocene je prisutna ukoliko koristimo nedovoljno dobar ocenjivač. Primer za ovaj slučaj je ocenjivač koji računa prosečnu vrednost visine tri najniža košarkaša.

Disperzija uzoračke raspodele od $\hat{t}$ je

$$V(\hat{t}) = E((\hat{t} - E(\hat{t}))^2) = \sum_{\mathcal{S}} P(\mathcal{S})(\hat{t}_{\mathcal{S}} - E(\hat{t}))^2,$$

gde se sumiranje vrši po svim mogućim uzorcima $\mathcal{S}$.

S obzirom da nekada koristimo pristrasne ocenjivače, za meru tačnosti ocenjivača se često koristi srednje kvadratna greška (eng. Mean Square Error - MSE)

$$MSE(\hat{t}) = E((\hat{t} - t)^2) = E((\hat{t} - E(\hat{t}) + E(\hat{t}) - t)^2)$$
$$= E((\hat{t} - E(\hat{t}))^2) + (E(\hat{t}) - t)^2 + 2E((\hat{t} - E(\hat{t}))(E(\hat{t}) - t))$$
$$= V(\hat{t}) + (Bias(\hat{t}))^2.$$

Možemo zaključiti da je ocenjivač $\hat{t}$:

- nepristrasan, ako je $Bias(\hat{t}) = 0$, odnosno $E(\hat{t}) = t$
- precizan, ako je $V(\hat{t}) = E((\hat{t} - E(\hat{t}))^2)$ malo
- tačan, ako je $MSE(\hat{t}) = E((\hat{t} - t)^2)$ malo.

Loš pristrasan ocenjivač može biti precizan, ali neće biti tačan. Tačnost (MSE) pokazuje koliko je ocena blizu tačne vrednosti parametra populacije, dok preciznost (disperzija) meri koliko su ocene dobijene iz različitih uzoraka bliske jedna drugoj.

Dakle, neka posmatrana populacija ima N jedinica $\{1, 2, ..., N\}$ i neka su $\{y_1, y_2, ..., y_N\}$ vrednosti posmatranog obeležja na tim jedinicama. Iz populacije biramo uzorak $\mathcal{S}$ obima n koristeći verovatnoće izbora. Vrednosti y_i su fiksne, ali nepoznate sve dok se jedinica ne nađe u našem uzorku $\mathcal{S}$. Vrednosti y_i, $i \in \mathcal{S}$ iz uzorka su jedina informacija koju posedujemo u vezi sa vrednostima obeležja cele populacije.

Populacione veličine koje se najčešće ispituju su:

- ukupna vrednost populacije (eng. population total)

$$t = \sum_{i=1}^{N} y_i$$

- srednja vrednost populacije (eng. mean)

$$\bar{y}_U = \frac{1}{N} \sum_{i=1}^{N} y_i.$$

U gotovo svim populacijama su prisutne različitosti. Na primer, različita domaćinstva imaju različite prihode, dok košarkaši reprezentacije imaju različite visine. Iz tog razloga, definišemo disperziju vrednosti populacije oko srednje vrednosti

$$S^2 = \frac{1}{N-1} \sum_{i=1}^{N} (y_i - \bar{y}_U)^2 \tag{1.1}$$

i standardnu devijaciju populacije

$$S = \sqrt{S^2}.$$

Koeficijent disperzije je bezdimenziona veličina koja meri relativno odstupanje

$$CV(y) = \frac{S}{\bar{y}_U}, \quad \bar{y}_U \neq 0.$$

Uvodimo još jednu veličinu, a to je proporcija jedinica populacije koje poseduju određenu karakteristiku.

Proporcija p je

$$p = \frac{broj\ \ jedinica\ \ sa\ \ datom\ \ karakteristikom\ \ u\ \ populaciji}{N}.$$

Proporcija jedinica koje imaju određenu karakteristiku u populaciji je samo specijalan slučaj sredine, koji se dobija stavljanjem $y_i = 1$, ako i-ta jedinica ima posmatranu karakteristiku, a $y_i = 0$, ako je i-ta jedinica nema.

2

Prost slučajan uzorak

Prosto slučajno uzorkovanje predstavlja osnovni oblik verovatnosnog uzorkovanja koji daje teorijsku osnovu za složenije oblike. Postoje dve vrste prostog slučajnog uzorka:

- sa ponavljanjem – u kojem svaka jedinica može biti uključena u uzorak više puta
- bez ponavljanja – u kojem su sve jedinice u uzorku različite.

Prost slučajan uzorak sa ponavljanjem veličine n iz populacije od N elemenata se može posmatrati kao uzimanje n nezavisnih uzoraka veličine 1. Međutim, u konačnim populacijama, biranje iste jedinice više puta ne pruža nikakvu dodatnu informaciju. Iz tog razloga se obično preferira prost slučajan uzorak bez ponavljanja, kako uzorak ne bi sadržao duplikate jedinica.

Prost slučajan uzorak bez ponavljanja (SRS) veličine n se bira tako da svaki mogući podskup od n različitih jedinica populacije ima jednaku verovatnoću da bude izabran za uzorak. U populaciji od N jedinica postoji $\binom{N}{n}$ mogućih uzoraka veličine n i svaki je jednako verovatan, pa je verovatnoća odabira proizvoljnog uzorka $\mathcal{S}$ jednaka

$$P(\mathcal{S}) = \frac{1}{\binom{N}{n}} = \frac{n!(N-n)!}{N!}.$$

Iz tog razloga je verovatnoća izbora svake jedinice

$$\pi_i = \sum_{i \in \mathcal{S}} P(\mathcal{S}) = \binom{N-1}{n-1} \frac{n!(N-n)!}{N!} = \frac{n}{N}.$$

Za uzimanje prostog slučajnog uzorka potrebno je imati listu svih jedinica populacije. Ta lista predstavlja uzorački okvir. U prostom slučajnom uzorku

se uzoračka jedinica i jedinica posmatranja poklapaju. Svakoj jedinici se pridružuje broj, a uzorak se bira tako da:

1. svaka jedinica ima istu šansu da se nađe u uzorku

2. na odabir jedinica ne utiču prethodno odabrane jedinice.

Ilustracija SRS predstavlja izvlačenje brojeva iz šešira. U praksi se za odabir uzorka koriste kompjuterski generisani pseudo slučajni brojevi.

Kao što je pomenuto, populacione veličine su obično nepoznate, pa se one ocenjuju na osnovu uzorka. Jedna od najčešće ispitivanih populacionih veličina je srednja vrednost populacije $\bar{y}_U$. Ocenjivač srednje vrednosti populacije $\bar{y}_U$ u SRS je uzoračka sredina $\bar{y}_{\mathcal{S}}$, oblika

$$\bar{y}_{\mathcal{S}} = \frac{1}{n} \sum_{i \in \mathcal{S}} y_i.$$

U daljem tekstu, za uzoračku sredinu ćemo koristiti oznaku $\bar{y}$, izostavljajući $\mathcal{S}$ u indeksu ako je jasno o kom uzorku je reč. Uzoračka sredina $\bar{y}$ je nepristrasan ocenjivač srednje vrednosti populacije $\bar{y}_U$. Disperzija od $\bar{y}$ je

$$V(\bar{y}) = \frac{S^2}{n}(1 - \frac{n}{N}),$$

gde je S^2 ranije definisana disperzija vrednosti populacije oko sredine, data sa (1.1). Disperzija $V(\bar{y})$ meri varijabilnost ocena $\bar{y}_U$ dobijenih iz različitih uzoraka.

Faktor $(1 - \frac{n}{N})$ se naziva populaciona korekcija (eng. finite population correction - fpc). Intuitivno, ova korekcija se vrši zbog malih populacija kod kojih sa povećanjem uzoračkog količnika $\frac{n}{N}$ imamo više informacija o populaciji i usled toga je disperzija manja. Za većinu uzoraka uzetih iz ekstremno velikih populacija, fpc je približno jednak 1. Prema tome, za velike populacije veličina uzorka određuje preciznost ocenjivača, a ne procenat uzorkovane populacije. Na primer, ukoliko je supa dobro skuvana i promešana, dovoljno je da probamo jednu ili dve kašičice da bismo odredili da li je ona dovoljno slana, bez obzira da li smo napravili 1 ili 20 litara supe.

Primer 2.1 *Uzorak veličine 100 iz populacije od 100000 elemenata ima gotovo jednaku preciznost kao uzorak veličine 100 uzet iz populacije od 100 miliona jedinica:*

- *za N=100000:* $V(\bar{y}) = \frac{S^2}{100}\frac{99900}{100000} = \frac{S^2}{100}(0.999)$
- *za N=100000000:* $V(\bar{y}) = \frac{S^2}{100}\frac{99999900}{1000000000} = \frac{S^2}{100}(0.999999)$ □

Pošto je disperzija populacije S^2 nepoznata jer zavisi od vrednosti čitave populacije, ona se ocenjuje uzoračkom disperzijom

$$s^2 = \frac{1}{n-1}\sum_{i\in\mathcal{S}}(y_i - \bar{y})^2.$$

Nepristrasan ocenjivač za disperziju od $\bar{y}$, tj. za $V(\bar{y})$ je

$$\hat{V}(\bar{y}) = (1 - \frac{n}{N})\frac{s^2}{n}.$$

Umesto ocenjene disperzije $\hat{V}(\bar{y})$, češće se koristi njen kvadratni koren koji se naziva standardna greška (eng. standard error - SE)

$$SE(\bar{y}) = \sqrt{(1 - \frac{n}{N})\frac{s^2}{n}}.$$

Prethodni rezultati mogu da se primene na ocenjivanje ukupne vrednosti populacije t, jer je

$$t = \sum_{i=1}^{N} y_i = N\bar{y}_U.$$

Za ocenu ukupne vrednosti populacije t koristimo nepristrasan ocenjivač

$$\hat{t} = N\bar{y}.$$

Tada je

$$V(\hat{t}) = N^2 V(\bar{y}) = N^2(1 - \frac{n}{N})\frac{S^2}{n}$$

i

$$\hat{V}(\hat{t}) = N^2(1 - \frac{n}{N})\frac{s^2}{n}.$$

Pored navedenih nepristrasnih ocenjivača ukupne i srednje vrednosti populacije, postoje i pristrasni ocenjivači, i o njima će biti reči u narednom poglavlju.

Od sad pa nadalje, pod pojmom ocenjivač podrazumeva se posmatrana slučajna promenljiva, dok je ocena konkretna realizacija te slučajne promenljive.

Intervali poverenja

Nakon odabira uzorka i ocenjivanja srednje i ukupne vrednosti populacije, poželjno je proceniti tačnost dobijenih ocena. To se obično postiže konstruisanjem intervala poverenja u okviru kojih se sa dovoljnom sigurnošću nalaze populacione veličine (parametri).

Neka I predstavlja interval poverenja za parametar θ. Ako je α verovatnoća greške, tada za interval poverenja važi

$$P(\theta \in I) = 1 - \alpha.$$

Krajevi intervala su statistike koje se menjaju od uzorka do uzorka, dok je parametar θ koji predstavlja tačnu vrednost populacione veličine obično nepoznat, ali fiksiran.

Veličina $1 - \alpha$ naziva se nivo poverenja, a interval I je $100(1 - \alpha)\%$-tni interval poverenja. Obično se za α biraju vrednosti 0.01, 0.05 i 0.1. Za $\alpha = 0.05$ nivo poverenja je 0.95. Kod prostog slučajnog uzorka, 95%-tni interval poverenja znači da interval pokriva tačnu vrednost parametra θ za 95% mogućih uzoraka veličine n. Drugim rečima, mi ne znamo da li 95%-tni interval poverenja koji smo konstruisali na osnovu uzorka sadrži tačnu vrednost parametra. Međutim, znamo da ako je procedura vezana za interval poverenja dobro sprovedena i ako ponavljamo proceduru dovoljan broj puta, možemo očekivati da 95% rezultujućih intervala poverenja sadrži tačnu vrednost parametra.

Približan $100(1-\alpha)\%$-tni interval poverenja za srednju i ukupnu vrednost populacije zasniva se na normalnoj aproksimaciji za raspodelu sredine uzorka kod prostog slučajnog uzorka.

U opštem slučaju, ako je $\hat{\theta}$ normalno raspoređena nepristrasna ocena za parametar populacije θ, tada je interval poverenja za θ sa nivoom poverenja $1 - \alpha$ dat sa

$$\hat{\theta} \pm z\sqrt{V(\hat{\theta})},$$

gde je z vrednost iz tablica za $\mathcal{N}(0, 1)$ raspodelu, takva da je

$$P\{|Z| \leq z\} = 1 - \alpha,$$

pri čemu Z ima normalnu $\mathcal{N}(0, 1)$ raspodelu. U praksi, $\hat{\theta}$ može imati raspodelu koja je približno normalna čak i kada date y-vrednosti nemaju normalnu raspodelu.

Pošto se disperzija ocenjivača obično ocenjuje iz uzorka, interval poverenja je oblika

$$\hat{\theta} \pm z\sqrt{\hat{V}(\hat{\theta})}.$$

Za 95%-tni interval poverenja, vrednost z iz normalne $\mathcal{N}(0,1)$ raspodele je 1.96. Dakle, za uzorak dovoljno velikog oblima ($n > 30$), 95%-tni interval poverenja je oblika

$$\hat{\theta} \pm 1.96 SE(\hat{\theta}),$$

pri čemu je $\hat{\theta}$ ocena parametra θ, a $SE(\hat{\theta})$ je standardna greška ocene.

Prost slučajan uzorak je osnovni oblik verovatnosnog uzorka koji je jednostavan za osmišljavanje i analizu. Zgodno ga je koristiti kada su prilikom dizajniranja istraživanja dostupni samo osnovni podaci, tj. kada je malo dodatnih informacija na raspolaganju. Na primer, ukoliko se uzorački okvir sastoji samo od spiska imena i prezimena studenata jednog univerziteta poređanih po azbučnom redosledu, bez dodatnih informacija o godini studija ili departmanu, prost slučajan uzorak je dobar izbor.

Osim toga, SRS treba koristiti kada osobe koje vrše analizu podataka insistiraju baš na upotrebi formula prostog slučajnog uzorka, bez obzira da li su one odgovarajuće ili ne. Na primer, prost slučajan uzorak se često preporučuje prilikom prilaganja uzoračkih dokaza u pravnim postupcima, jer ukoliko se koristi složenija uzoračka procedura, suprotstavljeni advokat može pokušati da ubedi porotu da rezultati dobijeni uzorkovanjem nisu validni.

3

Ocena količnika

Da bi se mogla primeniti ocena količnika potrebno je na svakoj uzoračkoj jedinici meriti veličine x_i i y_i, pri čemu se x_i često naziva pomoćna promenljiva. U populaciji veličine N tada je

$$t_y = \sum_{i=1}^{N} y_i,$$

$$t_x = \sum_{i=1}^{N} x_i,$$

dok je količnik ove dve veličine

$$B = \frac{t_y}{t_x} = \frac{\bar{y}_U}{\bar{x}_U}.$$

Najjednostavnije korišćenje ocene količnika sastoji se iz uzimanja prostog slučajnog uzorka veličine n i merenja veličina x i y u uzorku, koje se zatim koriste za ocenjivanje B, t_y ili $\bar{y}_U$.

Ocena količnika zavisi od korelacije između veličina x i y u populaciji i što je veća korelacija, bolje su ocene. Definišemo populacioni koeficijent korelacije R za x i y na sledeći način

$$R = \frac{\sum_{i=1}^{N}(x_i - \bar{x}_U)(y_i - \bar{y}_U)}{(N-1)S_x S_y},$$

pri čemu je S_x populaciona standardna devijacija za x_i, S_y je populaciona standardna devijacija za y_i, a R je Pearson-ov koeficijent korelacije između x i y za N jedinica populacije.

Primer 3.1 *Pretpostavimo da populaciju čine poljoprivredna polja različitih veličina. Neka su*

- y_i – *kilogrami žita u i-tom polju*
- x_i – *površina i-tog polja u hektarima.*

Tada je

- B – *prosečan prinos u kilogramima po hektaru*
- $\bar{y}_U$ – *prosečan prinos u kilogramima po polju*
- t_y – *ukupan prinos u kilogramima.* □

U slučaju da uzimamo prost slučajan uzorak, ocenjivači za B, t_y i $\bar{y}_U$ su:

$$\hat{B} = \frac{\bar{y}}{\bar{x}} = \frac{\hat{t}_y}{\hat{t}_x},$$

$$\hat{t}_{yr} = \hat{B}t_x,$$

$$\hat{\bar{y}}_r = \hat{B}\bar{x}_U,$$

gde se pretpostavlja da su t_x i $\bar{x}_U$ poznate veličine.

3.1 Razlozi za korišćenje ocene količnika

Postoji nekoliko razloga za korišćenje ocene količnika i oni će biti navedeni u ovom odeljku.

1. U nekim slučajevima upravo je količnik dve veličine, veličina koju želimo da ocenimo.

   U prethodnom primeru, prosečan prinos kilograma po hektaru B je ocenjen pomoću količnika uzoračkih sredina $\hat{B} = \bar{y}/\bar{x}$. Ukoliko su polja različitih veličina, i brojilac i imenilac su slučajne veličine. Ako izaberemo drugačiji uzorak, za očekivati je da će se i $\bar{x}$ i $\bar{y}$ promeniti.
   Neke ocene količnika su prikrivene jer imenilac izgleda kao da predstavlja veličinu uzorka. Međutim, ukoliko uzimanjem drugačijeg uzorka imenilac menja vrednost, tada treba koristiti ocenu količnika.

   Pretpostavimo da nas zanima procenat stranica jednog časopisa koje sadrže najmanje jednu reklamu. Možemo uzeti prost slučajan uzorak od deset izdanja tog časopisa i za svako izdanje dobiti sledeće podatke:

x_i – ukupan broj stranica u i-tom izdanju

y_i – ukupan broj stranica u i-tom izdanju koje sadrže najmanje jednu reklamu.

Proporcija koja nas u ovom slučaju zanima se ocenjuje na sledeći način

$$\hat{B} = \frac{\sum_{i \in S} y_i}{\sum_{i \in S} x_i}.$$

Imenilac predstavlja ukupan broj stranica u deset izdanja časopisa i za očekivati je da će biti različit za različite uzorke.

Možemo zaključiti da ocenu količnika koristimo svaki put kada uzimamo prost slučajan uzorak i ocenjujemo srednju vrednost ili proporciju za podpopulaciju.

2. Ponekad želimo da ocenimo ukupnu vrednost populacije, ali je veličina populacije N nepoznata, pa zato ne možemo da koristimo ocenjivač $\hat{t}_y = N\bar{y}$. Pošto znamo da je $N = t_x/\bar{x}_U$, a t_x i $\bar{x}_U$ su poznati, tada možemo da ocenimo N sa $t_x/\bar{x}$. Iz tog razloga, umesto veličine populacije N, koristimo drugu meru za veličinu, t_x.

   Za ocenjivanje ukupnog broja riba u ulovu čija je dužina veća od $20cm$, možemo uzeti prost slučajan uzorak riba, oceniti proporciju riba dužih od $20cm$ i pomnožiti je sa ukupnim brojem riba N. Ovaj račun ne možemo primeniti ukoliko je broj riba u ulovu N, nepoznat. Možemo, međutim, izmeriti težinu ulova ribe t_x, iz uzorka izračunati prosečnu težinu jedne ribe $\bar{x}$ i proporciju riba dužih od $20cm$, pa te podatke iskoristiti za procenu ukupnog broja riba dužih od $20cm$ u ulovu, pa je

   $$\hat{t}_{yr} = \bar{y}\frac{t_x}{\bar{x}}.$$

   Ukupna težina ulova t_x se jednostavno meri, dok $t_x/\bar{x}$ ocenjuje ukupan broj riba u ulovu.

3. Ocena količnika se često koristi u cilju povećanja preciznosti ocena ukupne i srednje vrednosti populacije.

   Laplace je koristio ocene količnika upravo iz ovog razloga. On je želeo da oceni ukupnu populaciju Francuske 1802. godine. Prvo je izabrao 30 opština i ustanovio broj stanovnika koji su u njima živeli. Zatim je posmatrao broj registrovanih rođenja u tim opštinama u periodu od 3 godine i došao do prosečnog godišnjeg broja rođenja u izabranih 30 opština. Deleći prosečan godišnji broj rođenja sa brojem stanovnika u tim opštinama, ocenio je da se svake godine, na

28.352845 stanovnika rodi jedna osoba. Rezonujući da veće opštine imaju veći broj registrovanih rođenja i pretpostavljajući da će odnos između broja stanovnika i godišnjih rođenja biti u čitavoj državi približno jednak odnosu u njegovom uzorku, zaključio je da se populacija Francuske može oceniti množenjem ukupnog broja godišnjih registrovanih rođenja u Francuskoj sa 28.352845. Laplace je ukupan broj registrovanih rođenja koristio samo kao pomoćnu promenljivu za ocenjivanje ukupne populacije Francuske. U ovom primeru je

y_i – broj osoba u i-toj opštini

x_i – broj registrovanih rođenja u i-toj opštini.

Laplace je mogao oceniti ukupnu populaciju Francuske množenjem prosečnog broja ljudi u 30 opština $\bar{y}$, sa ukupnim brojem opština u Francuskoj N. Međutim, on je pretpostavio da će korišćenjem ocene količnika postići veću preciznost. Što je veća populacija jedne opštine, za očekivati je da će biti veći i broj registrovanih rođenja. Iz tog razloga, logično je da će populacioni koeficijent korelacije R biti pozitivan, tj. $\bar{y}$ i $\bar{x}$ su pozitivno korelisani, pa će uzoračka raspodela za $\bar{y}/\bar{x}$ imati manju varijabilnost od uzoračke raspodele za $\bar{y}/\bar{x}_U$. Prema tome, ukoliko je nepoznat ukupan broj registrovanih rođenja t_x, za očekivati je da će srednje kvadratna greška od $\hat{t}_{yr} = \hat{B}t_x$ biti manja od srednje kvadratne greške od $N\bar{y}$, što znači da je ocena količnika preciznija od nepristrasne ocene.

4. Ocena količnika se koristi za prilagođavanje ocena iz uzorka tako da one oslikavaju ukupne demografske vrednosti.

   Na primer, prost slučajan uzorak od 400 studenata uzet sa univerziteta koji broji 4000 studenata može sadržati 240 ženskih i 160 muških osoba, od kojih 84 žene i 40 muškaraca planiraju da se zaposle u prosveti. Koristeći isključivo informacije iz prostog slučajnog uzorka, ocenjujemo da

   $$(\frac{4000}{400})124 = 1240$$

   studenata planiraju da se zaposle u prosveti. Znajući da univerzitet pohađa 2700 žena i 1300 muškaraca, može se dobiti i bolja ocena

   $$(\frac{84}{240})2700 + (\frac{40}{160})1300 = 1270.$$

   Ovde se ocena količnika koristi unutar svakog pola. 60% uzorka čine žene, ali one predstavljaju 67% populacije, pa prema tome prilagođava-

mo ocenu ukupnog broja studenata koji planiraju karijeru nastavnika. Za ocenjivanje ukupnog broja žena koje planiraju da postanu nastavnice koristimo

$$y_i = \begin{cases} 1 & \text{, ukoliko je žena i planira da postane nastavnica} \\ 0 & \text{, u suprotnom} \end{cases}$$

$$x_i = \begin{cases} 1 & \text{, ukoliko je žena} \\ 0 & \text{, u suprotnom} \end{cases}$$

Tada je

$$\frac{\sum_{i\in S} y_i}{\sum_{i\in S} x_i} t_x = (\frac{84}{240})2700 = 945$$

ocena količnika ukupnog broja žena koje planiraju karijeru u nastavi. Analogno

$$(\frac{40}{160})1300 = 325$$

je ocena količnika ukupnog broja muškaraca koji žele da postanu nastavnici.

Primer 3.2 *Vlada Sjedinjenih Američkih Država na svakih pet godina sprovodi popis poljoprivrede, prikupljajući podatke o svim farmama u svojih 50 država. Zahvaljujući popisu poljoprivrede moguć je uvid u podatke o ukupnom broju farmi, ukupnoj površini farmi, prinosu različitih useva, kao i o mnogim drugim poljoprivrednim merama za svaki od* $N = 3078$ *okruga u Sjedinjenim Državama. Pretpostavimo da, na osnovu popisa, imamo kompletne informacije o celoj populaciji za 1987. godinu i prost slučajan uzorak od 300 okruga iz 1992. godine od ukupnog broja okruga* $N = 3078$ *u Americi. Kada se ista veličina meri u različito vreme, podaci iz ranijeg merenja su obično odlična pomoćna promenljiva. Neka je:*

y_i *– ukupan broj ari farmi u i-tom okrugu u 1992. godini*

x_i *– ukupan broj ari farmi u i-tom okrugu u 1987. godini.*

Na osnovu popisa iz 1987. godine poznato je da je ukupan broj ari farmi u celoj Americi

$$t_x = 964470625,$$

pa je prosečan broj ari po okrugu

$$\bar{x}_U = \frac{964470625}{3078} = 313343.3.$$

Na osnovu podataka dobijenih iz uzorka 1992. godine sledi da je

$$\bar{y} = 297897.0,$$

$$\bar{x} = 301953.7,$$

pa je

$$\hat{B} = \frac{\bar{y}}{\bar{x}} = 0.986565$$

$$\hat{\bar{y}}_r = \hat{B}\bar{x}_U = 309133.6$$

$$\hat{t}_{yr} = \hat{B}t_x = 951513191.$$

□

3.2 Pristrasnost i srednje kvadratna greška ocenjivača količnika

Za razliku od nepristrasnog ocenjivača srednje vrednosti populacije $\bar{y}$ i nepristrasnog ocenjivača ukupne vrednosti populacije $\hat{t}_y = N\bar{y}$ kod prostog slučajnog uzorka, ocenjivači količnika za ocenjivanje $\bar{y}_U$ i t_y su obično pristrasni. Najpre posmatramo nepristrasnu ocenu $\bar{y}$. Ukoliko izračunamo $\bar{y}_{\mathcal{S}}$ za sve moguće proste slučajne uzorke $\mathcal{S}$, tada je prosek svih uzoračkih sredina od svih mogućih uzoraka baš srednja vrednost populacije $\bar{y}_U$. Međutim, pristrasnost ocene kod ocenjivanja količnika se javlja usled množenja $\bar{y}$ sa $\bar{x}_U/\bar{x}$. Ako izračunamo $\hat{\bar{y}}_r$ za sve moguće proste slučajne uzorke $\mathcal{S}$, prosek svih vrednosti od $\hat{\bar{y}}_r$ iz različitih uzoraka će biti približno, ali ne nužno jednak $\bar{y}_U$, što znači da je ocenjivač količnika $\hat{\bar{y}}_r$ pristrasan.

Manja disperzija ocenjivača količnika obično kompenzuje prisustvo pristrasnosti. Iako je $\hat{\bar{y}}_r$ pristrasan ocenjivač, tj. $E(\hat{\bar{y}}_r) \neq \bar{y}_U$, vrednost $\hat{\bar{y}}_r$ za proizvoljan uzorak je verovatno bliža tačnoj vrednosti $\bar{y}_U$ nego što je uzoračka sredina $\bar{y}_{\mathcal{S}}$. Pošto se u praksi uzima samo jedan uzorak, obično se ističe da je ocena količnika dobijena iz uzorka veoma bliska tačnoj vrednosti. Sa druge strane, ocena dobijena na osnovu nepristrasnog ocenjivača $\bar{y}_{\mathcal{S}}$ može biti prilično daleko od $\bar{y}_U$, ali je odstupanje $\bar{y}_{\mathcal{S}} - \bar{y}_U$ usrednjeno kroz sve moguće uzorke jednako 0. Ocenjivač količnika je pristrasan, ali je često precizniji od nepristrasnog ocenjivača.

Prilikom izračunavanja pristrasnosti i disperzije ocenjivača količnika koristi se sledeći identitet

$$\hat{t}_{yr} - t_y = \frac{\hat{t}_y}{\hat{t}_x}t_x - t_y = \hat{t}_y(1 - \frac{\hat{t}_x - t_x}{\hat{t}_x}) - t_y.$$

Pošto je $E(\hat{t}_y) = t_y$, sledi

$$E(\hat{t}_{yr} - t_y) = E(\hat{t}_y) - t_y - E(\frac{\hat{t}_y}{\hat{t}_x}(\hat{t}_x - t_x))$$

$$= -E(\hat{B}(\hat{t}_x - t_x))$$
$$= -Cov(\hat{B}, \hat{t}_x)$$

i zaključujemo da je

$$E(\hat{B} - B) = \frac{E(\hat{t}_{yr} - t_y)}{t_x} = \frac{-Cov(\hat{B}, \bar{x})}{\bar{x}_U}.$$

Zbog toga je

$$\frac{\mid Bias(\hat{B}) \mid}{[V(\hat{B})]^{1/2}} = \mid \frac{Corr(\hat{B}, \bar{x})}{\bar{x}_U} \mid (\frac{V(\hat{B})V(\bar{x})}{V(\hat{B})})^{1/2} \leq \frac{V(\bar{x})^{1/2}}{\bar{x}_U} = CV(\bar{x}).$$

Prema tome, kod prostog slučajnog uzorka, apsolutna vrednost pristrasnosti ocenjivača količnika je mala u poređenju sa standardnom devijacijom ocenjivača, ukoliko je $CV(\bar{x})$ malo.

Takođe, može se pokazati da je

$$E(\hat{B} - B) \approx (1 - \frac{n}{N})\frac{1}{n\bar{x}_U^2}(BS_x^2 - RS_xS_y)$$

$$= \frac{1}{\bar{x}_U^2}[BV(\bar{x}) - Cov(\bar{x}, \bar{y})],$$

gde je R korelacija između x i y.
Dakle, pristrasnost ocenjivača $\hat{B}$ je mala ako je:

- veličina uzorka n velika
- $\frac{n}{N}$ veliko
- $\bar{x}_U$ veliko
- S_x malo
- R blizu 1.

Za ocenu srednje kvadratne greške od $\hat{B}$ dobija se

$$E((\hat{B}-B)^2) = E((\frac{\bar{y}-B\bar{x}}{\bar{x}})^2) = E((\frac{\bar{y}-B\bar{x}}{\bar{x}_U}(1-\frac{\bar{x}-\bar{x}_U}{\bar{x}}))^2)$$

$$= E((\frac{\bar{y}-B\bar{x}}{\bar{x}_U})^2 + (\frac{\bar{y}-B\bar{x}}{\bar{x}_U})^2((\frac{\bar{x}-\bar{x}_U}{\bar{x}})^2 - 2\frac{\bar{x}-\bar{x}_U}{\bar{x}})).$$

Imenilac u prvom članu je konstanta, a ne slučajna promenljiva. Može se pokazati da je drugi član mali u poređenju sa prvim, pa se MSE može aproksimirati sa

$$E((\hat{B}-B)^2) \approx E((\frac{\bar{y}-B\bar{x}}{\bar{x}_U})^2) = \frac{1}{\bar{x}_U^2}E((\bar{y}-B\bar{x})^2).$$

Sada definišemo

$$d_i = y_i - Bx_i.$$

Tada je $\bar{d} = \bar{y} - B\bar{x}$, $E(\bar{d}) = 0$, pa je

$$E((\bar{y}-B\bar{x})^2) = V(\bar{d}) = (1-\frac{n}{N})\frac{1}{n}\sum_{i=1}^{N}\frac{d_i^2}{N-1} \tag{3.1}$$

i

$$E((\hat{B}-B)^2) \approx \frac{1}{\bar{x}_U^2}V(\bar{d}).$$

Jednakost (3.1) je ekvivalentna sa

$$\frac{1}{\bar{x}_U^2}E((\bar{y}-B\bar{x})^2) = (1-\frac{n}{N})\frac{1}{n\bar{x}_U^2}(S_y^2 - 2BRS_xS_y + B^2S_x^2). \tag{3.2}$$

Iz (3.1) i (3.2) zaključujemo da će aproksimirana MSE biti mala kada je

- veličina uzorka n velika
- $\frac{n}{N}$ veliko
- $\bar{x}_U$ veliko
- devijacija oko prave $y = Bx$ mala
- R blizu 1.

U praksi je B nepoznato, pa ne možemo izračunati d_i za uzoračke vrednosti. Umesto toga, koristimo

$$e_i = y_i - \hat{B}x_i,$$

što je i-ti ostatak iz fitovanja prave $y = \hat{B}x$. Disperziju od $\hat{B}$ ocenjujemo sa

$$\hat{V}(\hat{B}) = (1 - \frac{n}{N})\frac{s_e^2}{n\bar{x}_U^2} = (1 - \frac{n}{N})\frac{1}{n\bar{x}_U^2}\frac{\sum_{i\in S}(y_i - \hat{B}x_i)^2}{n-1}. \tag{3.3}$$

Ako je $\bar{x}_U$ nepoznato, zamenjujemo ga sa $\bar{x}_{\mathcal{S}}$.
Sada dobijamo

$$\hat{V}(\hat{t}_{yr}) = \hat{V}(t_x\hat{B}) = N^2(1 - \frac{n}{N})\frac{s_e^2}{n}$$

i

$$\hat{V}(\hat{\bar{y}}_r) = \hat{V}(\bar{x}_U\hat{B}) = (1 - \frac{n}{N})\frac{s_e^2}{n}.$$

Ako su veličine uzorka dovoljno velike, 95%-tni intervali poverenja se konstruišu na sledeći način

$$\hat{B} \pm 1.96SE(\hat{B}),$$

$$\hat{\bar{y}}_r \pm 1.96SE(\hat{\bar{y}}_r),$$

$$\hat{t}_{yr} \pm 1.96SE(\hat{t}_{yr}).$$

U velikim uzorcima pristrasnost ocenjivača je obično mala u odnosu na standardnu grešku, pa se efekat pristrasnosti u intervalima poverenja može ignorisati.
Napomenimo još da ako su svi x jednaki ($S_x = 0$), tada je nepristrasan ocenjivač prostog slučajnog uzorkovanja isti kao ocenjivač količnika, tj. važi

$$\hat{\bar{y}}_r = \bar{y},$$

$$SE(\hat{\bar{y}}_r) = SE(\bar{y}).$$

Primer 3.3 *Vratimo se primeru 3.2 od ranije. Izračunaju se ostaci* $e_i = y_i - \hat{B}x_i$ *i odgovarajući* s_e. *Tada je*

$$SE(\hat{t}_{yr}) = \sqrt{\hat{V}(\hat{t}_{yr})} = 3078\sqrt{1 - \frac{300}{3078}}\frac{s_e}{\sqrt{300}} = 5344568.$$

95%-tni interval poverenja za ukupnu površinu farmi korišćenjem ocenjivača količnika je

$$951513191 \pm 1.96 \times 5344568 = (941037838, 961988544).$$

Korišćenjem nepristrasnog ocenjivača $\hat{t}_y = N\bar{y}_S$ dobijena je standardna greška koja je više od deset puta veća

$$SE(N\bar{y}_S) = 3078\sqrt{(1 - \frac{300}{3078})}\frac{s_y}{\sqrt{300}} = 58169381,$$

što znači da je ocenjivač količnika precizniji od nepristrasnog ocenjivača. Koeficijent disperzije dobijen pomoću ocenjivača količnika je $CV(\hat{t}_{yr}) = \frac{5344568}{951513191}$ $= 0.0056$*, dok je za nepristrasan ocenjivač koeficijent disperzije* 0.0634*, na osnovu čega se, takođe, zaključuje da su informacije iz 1987. godine uz korišćenje ocenjivača količnika značajno povećale preciznost.* □

3.2.1 Prednosti ocene količnika

Prirodno se postavlja pitanje šta dobijamo korišćenjem ocene količnika. Ako su odstupanja y_i od $\hat{B}x_i$ manja od odstupanja y_i od $\bar{y}$, tada je

$$\hat{V}(\hat{\bar{y}}_r) \leq \hat{V}(\bar{y}).$$

Pošto je $\bar{y}$ nepristrasan ocenjivač za $\bar{y}_U$ sledi

$$MSE(\bar{y}) = V(\bar{y}) = (1 - \frac{n}{N})\frac{S_y^2}{n}.$$

Koristeći (3.2) dobija se

$$MSE(\hat{\bar{y}}_r) \approx (1 - \frac{n}{N})\frac{1}{n}(S_y^2 - 2BRS_xS_y + B^2S_x^2).$$

Prema tome je

$$MSE(\hat{\bar{y}}_r) - MSE(\bar{y}) \approx (1 - \frac{n}{N})\frac{1}{n}(S_y^2 - 2BRS_xS_y + B^2S_x^2 - S_y^2)$$

$$= (1 - \frac{n}{N})\frac{1}{n}S_xB(-2RS_y + BS_x).$$

Može se zaključiti da je

$$MSE(\hat{\bar{y}}_r) \leq MSE(\bar{y})$$

ako i samo ako je

$$R \geq \frac{BS_x}{2S_y} = \frac{CV(x)}{2CV(y)}.$$

Ako su koeficijenti disperzije približno jednaki, tada se isplati koristiti ocenu količnika kada je korelacija između x i y veća od $\frac{1}{2}$. Ocena količnika je najpo-

godnija ako prava koja prolazi kroz koordinatni početak odražava odnos između x_i i y_i i ako je disperzija y_i oko te prave proporcionalna sa x_i. Pod ovim uslovima, $\hat{B}$ je težinski nagib prave kroz koordinatni početak, sa težinama proporcionalnim $\frac{1}{x_i}$, tj. nagib $\hat{B}$ minimizuje sumu kvadrata

$$\sum_{i \in S} \frac{1}{x_i}(y_i - \hat{B}x_i)^2.$$

Ostaje još da napomenemo da ocena količnika funkcioniše na isti način i kada radimo sa proporcijama.

4

Klaster uzorci sa jednakim verovatnoćama

Kod prostog slučajnog uzorkovanja, pretpostavlja se da je data populacija iz koje uzimamo odgovarajući uzorak jedinica. Međutim, može se desiti da populacione jedinice nisu uvek jasno definisane, čak i kada populacija jeste. Osim toga, spisak svih članova populacije može biti nepoznat.

Na primer, pretpostavimo da želimo odrediti broj biciklova u vlasništvu stanovnika jedne zajednice koja se sastoji od 10000 domaćinstava. Jedan način je da uzmemo prost slučajan uzorak od 400 domaćinstava. Drugi način je da zajednicu podelimo na blokove od po 20 domaćinstava i da uzorkujemo svako domaćinstvo (ili biramo poduzorak domaćinstava iz blokova i onda izvršimo uzorkovanje) u svakom od 20 slučajno izabranih blokova od ukupno 500 blokova u zajednici. Ovaj drugi način je primer klaster uzorka. Blokovi predstavljaju primarne uzoračke jedinice odnosno klastere. Domaćinstva su sekundarne uzoračke jedinice. Treba napomenuti da su najčešće sekundarne uzoračke jedinice upravo elementi populacije.

Za očekivati je da klaster uzorak od 400 domaćinstava daje manju preciznost od prostog slučajnog uzorka od 400 domaćinstava. U našem konkretnom primeru, razlog za to je sledeći: u nekim blokovima zajednice žive pretežno porodice, dok su drugi blokovi naseljeni uglavnom penzionerima, tako da će verovatno blokovi naseljeni porodicama posedovati veći broj biciklova od blokova u kojima žive penzioneri. 20 domaćinstava iz istog bloka verovatno neće odražavati raznolikost zajednice jednako dobro kao 20 slučajno odabranih domaćinstava. Iz tog razloga će klaster uzorak, u ovoj situaciji, verovatno pružiti manje informacija po opservaciji nego što bi pružio prost slučajan uzorak iste veličine. Međutim, mnogo je jeftinije i jednostavnije intervjuisati svih 20 domaćinstava u istom bloku, nego 20 slučajno izabranih

domaćinstava iz zajednice, tako da je klaster uzorkovanje jeftinije i jednostavnije.

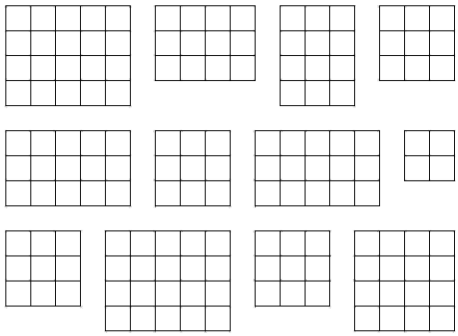
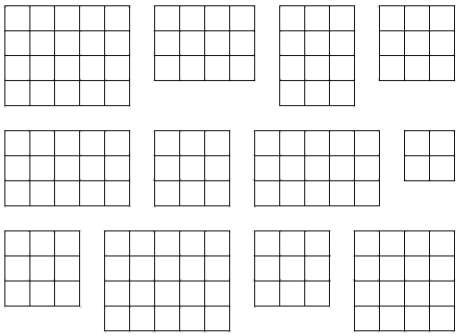
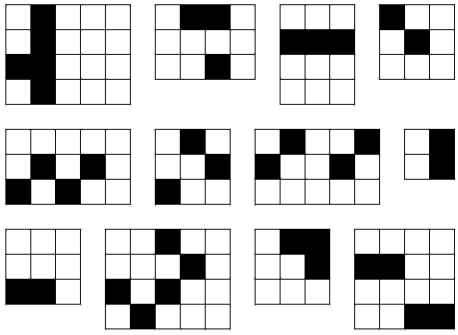
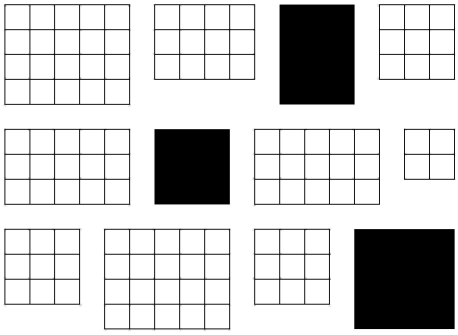
U klaster uzorku, pojedinačne jedinice populacije mogu biti u uzorku samo ukoliko pripadaju klasteru (primarnoj uzoračkoj jedinici) koji je uključen u uzorak. Uzoračke jedinice (primarne uzoračke jedinice) su klasteri i oni se razlikuju od jedinica posmatranja (sekundarnih uzoračkih jedinica). Iz tog razloga, pri izračunavanju standardnih grešaka klaster uzorka, u obzir se moraju uzeti dve veličine eksperimentalnih jedinica.

Razlozi za korišćenje klaster uzorka su sledeći:

1. Formiranje uzoračkog okvira može biti teško, skupo ili nemoguće. Na primer, ne možemo nabrojati sve pčele u određenom regionu ili sve kupce neke prodavnice. Možda možemo nabrojati svo drveće neke šume ili sve stanovnike grada za koji imamo podatke o svim domaćinstvima, ali formiranje ovakvih lista može biti veoma skupo i dugotrajno.

2. Populacija može biti geografski široko rasprostranjena ili može biti podeljena u prirodne klastere kao što su domaćinstva ili škole. Ukoliko ciljnu populaciju predstavljaju bolesnici koji leže u bolnicama, mnogo je jeftinije i jednostavnije uzorkovati bolnice i zatim intervjuisati sve bolesnike u izabranim bolnicama, nego intervjuisati bolesnike izabrane prostim slučajnim uzorkovanjem, jer bi se u tom slučaju moglo dogoditi da se mora putovati u određenu bolnicu radi intervjuisanja samo jednog bolesnika.

Klasteri imaju površnu sličnost sa stratumima jer klaster, kao i stratum, predstavlja grupu elemenata populacije. Međutim, suštinska razlika između klaster uzorka i stratifikovanog uzorka ogleda se u postupku selekcije tj. odabira. Sličnosti i razlike između ova dva tipa uzoraka ilustrovane su na sledećoj slici.

Stratifikovano uzorkovanje	Klaster uzorkovanje
Svaki element populacije se nalazi u tačno jednom stratumu.	Svaki element populacije se nalazi u tačno jednom klasteru.
Populacija od H stratuma: stratum h ima N_h elemenata 	Jednofazno klaster uzorkovanje: populacija od N klastera 
Uzima se SRS iz svakog stratuma: 	Uzima se SRS klastera i posmatraju se svi elementi unutar izabranih klastera: 
Disperzija ocene srednje vrednosti populacije $\bar{y}_U$ zavisi od promenljivosti unutar stratuma.	Klaster je uzoračka jedinica. Što više klastera uzorkujemo, manja je disperzija. Disperzija ocene srednje vrednosti populacije $\bar{y}_U$ zavisi prvenstveno od promenljivosti između sredina klastera.
Za postizanje veće preciznosti, pojedinačni elementi unutar svakog stratuma treba da budu što sličniji, a sredine stratuma treba da se razlikuju što je više moguće.	Za postizanje veće preciznosti, pojedinačni elementi unutar svakog klastera treba da budu što različitiji, dok sredine klastera treba da budu što sličnije.

U opštem slučaju, stratifikacija povećava preciznost u poređenju sa prostim slučajnim uzorkovanjem, a klaster uzorkovanje smanjuje preciznost. Članovi istog klastera imaju tendenciju da budu sličniji nego slučajno odabrani elementi iz cele populacije. Na primer, članovi jednog domaćinstva često imaju slična politička gledišta. Ribe u istom jezeru imaju slične koncentracije žive. Bolesnici jedne bolnice imaju slične utiske o kvalitetu nege pacijenata. Do ovih sličnosti dolazi usled nekih faktora koji mogu, a ne moraju biti merljivi. Bolesnici jedne bolnice mogu imati slična mišljenja jer je briga o pacijentima na niskom nivou, koncentracija žive u ribama će odražavati koncentraciju žive u jezeru... Zato, ispitivanjem dva bolesnika iste bolnice dobijamo manje informacija o svim bolesnicima jedne države nego ispitivanjem dva bolesnika u različitim bolnicama, jer su bolesnici iste bolnice sličniji. U klaster uzorku, ispitivanjem svih jedinica klastera, delimično ponavljamo istu informaciju, tj. dobijamo manje novih informacija nego što je slučaj kod SRS i to nam pruža manju preciznost ocena populacionih veličina.

Uprkos tome, klaster uzorkovanje se veoma često koristi u praksi jer je jeftinije i zgodnije populaciju uzorkovati pomoću klastera nego slučajnim odabirom. Gotovo sva veća ispitivanja domaćinstava vršena od strane vlade Sjedinjenih Američkih Država se sprovode pomoću klaster uzorkovanja zbog uštede troškova.

Sa druge strane, treba napomenuti da je jedna od najvećih grešaka prilikom klaster uzorkovanja, analiziranje podataka kao da su dobijeni prostim slučajnim uzorkovanjem. Usled ovakve greške, istraživači često dolaze do standardnih grešaka koje su znatno manje nego što bi realno trebalo da budu, što daje lažan utisak o većoj preciznosti istraživanja od stvarne.

Primer 4.1 *Basow i Silberg (1987) su radili istraživanje na temu da li studenti drugačije ocenjuju profesorke i profesore. Autori su uparili 16 profesora i 16 profesorki po zvanju, predmetu koji predaju i godinama nastavnog iskustva, a zatim su podelili evaluacione upitnike studentima u klasama tih profesora. Veličina uzorka je* $n = 32$ *i predstavlja broj posmatranih fakulteta, dok je 1029 studenata popunilo upitnike. Procene studenata odražavaju različite stilove predavanja. Za očekivati je da će procene studenata u okviru iste klase biti u većoj ili manjoj meri slične. One ne bi trebalo da se tretiraju kao nezavisne opservacije jer će verovatno biti pozitivno korelisane. Ukoliko se ova pozitivna korelacija zanemari, a procene tretiraju kao nezavisne, razlike će biti statistički značajnije nego što bi trebalo.* □

4.1 Oznake u klaster uzorkovanju

U prostom slučajnom uzorkovanju uzoračke jedinice su istovremeno i jedinice posmatranja. U klaster uzorkovanju, uzoračke jedinice su klasteri, tzv. primarne uzoračke jedinice, a jedinice posmatranja su sekundarne uzoračke jedinice i one su unutar klastera. Uvodimo sledeće oznake:

U – populacija od N primarnih uzoračkih jedinica

$\mathcal{S}$ – uzorak primarnih uzoračkih jedinica izabranih iz populacije U

$\mathcal{S}_i$ – uzorak sekundarnih uzoračkih jedinica izabranih iz i-te primarne uzoračke jedinice

y_{ij} – vrednost j-tog elementa u i-toj primarnoj uzoračkoj jedinici.

Primarnu uzoračku jedinicu ćemo u daljem tekstu obeležavati sa PSU (eng. Primary Sampling Unit), a sekundarnu uzoračku jedinicu sa SSU (eng. Secondary Sampling Unit).

Bez obzira na definisanje, notacija za klaster uzorkovanje je komplikovana jer su oznake potrebne i za nivo primarnih i za nivo sekundarnih uzoračkih jedinica. Treba napomenuti da N predstavlja broj primarnih uzoračkih jedinica, a ne broj posmatranih jedinica.

PSU nivo - populacione veličine

N – broj PSU u populaciji

M_i – broj SSU u i-toj PSU

$K = \sum_{i=1}^{N} M_i$ – ukupan broj SSU u populaciji

$t_i = \sum_{j=1}^{M_i} y_{ij}$ – ukupna vrednost populacije u i-toj PSU

$t = \sum_{i=1}^{N} t_i = \sum_{i=1}^{N} \sum_{j=1}^{M_i} y_{ij}$ – ukupna vrednost populacije

$S_t^2 = \sum_{i=1}^{N} \frac{(t_i - \frac{t}{N})^2}{N-1}$ – populaciona disperzija ukupnih vrednosti PSU

SSU nivo - populacione veličine

$\bar{y}_U = \sum_{i=1}^{N} \sum_{j=1}^{M_i} \frac{y_{ij}}{K}$ – srednja vrednost populacije

$\bar{y}_{iU} = \sum_{j=1}^{M_i} \frac{y_{ij}}{M_i} = \frac{t_i}{M_i}$ – srednja vrednost populacije u i-toj PSU

$S^2 = \sum_{i=1}^{N} \sum_{j=1}^{M_i} \frac{(y_{ij}-\bar{y}_U)^2}{K-1}$ – disperzija populacije (po SSU)

$S_i^2 = \sum_{j=1}^{M_i} \frac{(y_{ij}-\bar{y}_{iU})^2}{M_i-1}$ – disperzija populacije unutar i-te PSU

Uzoračke veličine

n – broj PSU u uzorku

m_i – broj elemenata u uzorku iz i-te PSU

$\bar{y}_i = \sum_{j \in \mathcal{S}_i} \frac{y_{ij}}{m_i}$ – uzoračka sredina (po SSU) za i-tu PSU

$\hat{t}_i = M_i \bar{y}_i = \sum_{j \in \mathcal{S}_i} \frac{M_i}{m_i} y_{ij}$ – ocenjivač ukupne vrednosti populacije za i-tu PSU

$\hat{t}_{unb} = \sum_{i \in \mathcal{S}} \frac{N}{n} \hat{t}_i$ – nepristrasan ocenjivač ukupne vrednosti populacije

$s_t^2 = \frac{1}{n-1} \sum_{i \in \mathcal{S}} (\hat{t}_i - \frac{\hat{t}_{unb}}{N})^2$ – ocenjena disperzija ukupne vrednosti PSU

$s_i^2 = \sum_{j \in \mathcal{S}_i} \frac{(y_{ij}-\bar{y}_i)^2}{m_i-1}$ – uzoračka disperzija unutar i-te PSU

4.2 Jednofazno klaster uzorkovanje

U jednofaznom klaster uzorkovanju ili se svi elementi klastera (primarne uzoračke jedinice) nalaze u uzorku ili se nijedan element klastera ne nalazi u uzorku. Jednofazno klaster uzorkovanje se koristi u mnogim istraživanjima kod kojih su troškovi uzorkovanja sekundarnih uzoračkih jedinica zanemarljivi u poređenju sa troškovima uzorkovanja primarnih uzoračkih jedinica. Na primer, u istraživanjima u obrazovanju, prirodni klasteri su razredi. Najčešće su svi učenici jednog razreda uključeni u istraživanje, s obzirom da je malo dodatnih troškova potrebno da bi se upitnik podelio svim učenicima umesto samo nekima.

U populaciji od N primarnih uzoračkih jedinica, i-ta primarna uzoračka jedinica sadrži M_i sekundarnih uzoračkih jedinica. Prvo biramo prost slučajan uzorak od n primarnih uzoračkih jedinica iz populacije, a zatim ispitujemo željeno obeležje na svakom elementu izabrane primarne uzoračke jedinice. Iz tog razloga, u jednofaznom klaster uzorku važi $M_i = m_i$.

4.2.1 Klasteri jednakih veličina - ocenjivanje

Najjednostavniji slučaj jednofaznog klaster uzorka jeste onaj u kom se svaki klaster sastoji iz jednakog broja elemenata, tj. $M_i = m_i = M$. Klasteri ljudi

obično ne pripadaju ovom tipu, ali se on javlja u poljoprivrednim i industrijskim istraživanjima. Ocenjivanje srednje i ukupne vrednosti populacije je jednostavno. Sredine klastera ili ukupne vrednosti klastera se tretiraju kao jedinice posmatranja, dok se pojedinačni elementi ignorišu.

Prema tome, imamo prost slučajan uzorak od n jedinica $\{t_i, i \in \mathcal{S}\}$, gde je t_i ukupna vrednost svih elemenata i-te primarne uzoračke jedinice. Tada, $\bar{t}_{\mathcal{S}}$ ocenjuje prosečnu ukupnu vrednost klastera. Na primer, u istraživanju domaćinstava za ocenjivanje prihoda dvočlanih domaćinstava, individualne jedinice posmatranja y_{ij} predstavljaju individualne prihode unutar domaćinstva, t_i je ukupan prihod i-tog domaćinstva (koji je poznat za domaćinstva koja se nalaze u uzorku jer su obe osobe intervjuisane), $\bar{t}_U$ je prosečan prihod po domaćinstvu, dok je $\bar{y}_U$ prosečan prihod po osobi. Za ocenu ukupnog prihoda t, koristimo ocenjivač

$$\hat{t} = \frac{N}{n} \sum_{i \in S} t_i.$$

Teorijski rezultati vezani za prost slučajan uzorak se mogu sada primeniti na $\hat{t}$ jer imamo SRS od n jedinica iz populacije od N jedinica. Zbog toga, zaključujemo da je $\hat{t}$ nepristrasan ocenjivač za t, sa disperzijom

$$V(\hat{t}) = N^2(1 - \frac{n}{N})\frac{S_t^2}{n}$$

i standardnom greškom

$$SE(\hat{t}) = N\sqrt{(1 - \frac{n}{N})\frac{s_t^2}{n}},$$

gde su S_t^2 i s_t^2, respektivno, populaciona i uzoračka disperzija ukupnih vrednosti PSU:

$$S_t^2 = \frac{1}{N-1} \sum_{i=1}^{N} (t_i - \frac{t}{N})^2$$

i

$$s_t^2 = \frac{1}{n-1} \sum_{i \in S} (t_i - \frac{\hat{t}}{N})^2.$$

Za ocenjivanje $\bar{y}_U$, treba podeliti ocenjenu ukupnu vrednost populacije sa brojem osoba, odakle sledi

$$\hat{\bar{y}} = \frac{\hat{t}}{NM},$$

sa disperzijom

$$V(\hat{\bar{y}}) = (1 - \frac{n}{N})\frac{S_t^2}{nM^2}$$

i standardnom greškom

$$SE(\hat{\bar{y}}) = \frac{1}{M}\sqrt{(1 - \frac{n}{N})\frac{s_t^2}{n}}.$$

Kod jednofaznog klaster uzorkovanja nema novina, već se samo koriste rezultati prostog slučajnog uzorkovanja kod kog su jedinice posmatranja zapravo ukupne vrednosti klastera.

Primer 4.2 *Student želi da oceni prosečnu ocenu svih studenata u svom domu. Umesto prikupljanja liste svih studenata i uzimanja prostog slučajnog uzorka, student je uočio da u domu postoji 100 apartmana i da u svakom apartmanu borave 4 studenta. Slučajnim odabirom je izabrao 5 apartmana u domu i upitao svakog studenta iz izabranih apartmana za prosečnu ocenu. Tom prilikom dobijeni su podaci navedeni u tabeli*

osoba	*apartman 1*	*apartman 2*	*apartman 3*	*apartman 4*	*apartman 5*
1	*3.08*	*2.36*	*2.00*	*3.00*	*2.68*
2	*2.60*	*3.04*	*2.56*	*2.88*	*1.92*
3	*3.44*	*3.28*	*2.52*	*3.44*	*3.28*
4	*3.04*	*2.68*	*1.88*	*3.64*	*3.20*
ukupno	*12.16*	*11.36*	*8.96*	*12.96*	*11.08*

PSU su apartmani, tako da je $N = 100$, $n = 5$ *i* $M = 4$. *Ocena ukupne vrednosti populacije (ocena sume svih prosečnih ocena za sve studente u domu), iako nebitna veličina za ovaj primer, ali korisna za ilustraciju procedure je*

$$\hat{t} = \frac{100}{5}(12.16 + 11.36 + 8.96 + 12.96 + 11.08) = 1130.4,$$

dok je

$$s_t^2 = \frac{1}{5-1}[(12.16 - 11.304)^2 + ... + (11.08 - 11.304)^2] = 2.256.$$

U ovom primeru, s_t^2 *je uobičajena uzoračka disperzija ukupnih vrednosti 5 apartmana. Prema tome, koristeći ranije formule dobijamo*

$$\hat{\bar{y}} = \frac{1130.4}{400} = 2.826,$$

$$SE(\hat{\bar{y}}) = \sqrt{(1 - \frac{5}{100})\frac{2.256}{(5)(4)^2}} = 0.164.$$

U navedenom računu korišćeni su samo podaci iz poslednje, ukupne, vrste tabele. Individualni proseci su korišćeni samo za dobijanje ukupnog proseka apartmana. □

Jednofazno klaster uzorkovanje, sa prostim slučajnim uzorkom primarnih uzoračkih jedinica, daje i samotežinski uzorak. Težina za svaku posmatranu jedinicu je

$$w_{ij} = \frac{1}{P\{j\text{-ta SSU iz } i\text{-te PSU je u uzorku}\}} = \frac{N}{n}.$$

Za podatke iz prethodnog primera sledi

$$\hat{t} = \sum_{i \in S} \sum_{j \in S_i} w_{ij} y_{ij} = \frac{N}{n}(3.08 + 2.60 + ... + 3.28 + 3.20) = \frac{100}{5}(56.52) = 1130.4.$$

Dakle, ukupnu vrednost populacije možemo oceniti sumiranjem proizvoda vrednosti posmatranih jedinica i uzoračkih težina.

Da smo uzeli prost slučajan uzorak od nM elemenata, svaki element u uzorku bi imao težinu $(NM)/(nM) = N/n$, što je ista težina koju smo dobili za klaster uzorak. Međutim, preciznost u ova dva tipa uzorkovanja može značajno da se razlikuje. Upravo o ovim razlikama u preciznosti govorimo u narednom odeljku.

4.2.2 Klasteri jednakih veličina - teorija

U ovom delu će biti poređeno klaster uzorkovanje i prosto slučajno uzorkovanje. Klaster uzorkovanje gotovo uvek daje manju preciznost ocenjivača od one koja se postiže prostim slučajnim uzorkovanjem sa istim brojem elemenata.
Uvodimo sledeće oznake:

SSB (eng. Sum of Squares Between PSU) – suma kvadrata između PSU

SSW (eng. Sum of Squares Within PSU) – suma kvadrata unutar PSU

SSTO (eng. Sum of Squares Total) – ukupna suma kvadrata

MSB (eng. Mean Square Between PSU) – sredina kvadrata između PSU

MSW (eng. Mean Square Within PSU) – sredina kvadrata unutar PSU

Df (eng. Degree of freedom) – broj stepeni slobode

Posmatrajmo ANOVA (eng. Analysis of Variance) tabelu za celu populaciju. U stratifikovanom uzorkovanju, disperzija ocenjivača od t zavisi od promenljivosti unutar stratuma i mala je ukoliko je SSW malo u odnosu na SSTO ili, ekvivalentno, ako je MSW malo u odnosu na S^2. U stratifikovanom uzorkovanju posedujemo informacije o svakom stratumu zbog čega nemamo razloga za brigu o promenljivosti usled neuzorkovanih stratuma. Ukoliko je količnik MSB/MSW velik, tj. promenljivost među sredinama stratuma je velika u poređenju sa promenljivostima unutar stratuma, onda stratifikovano uzorkovanje povećava preciznost.

Izvor	Df	Suma kvadrata	Sredina kvadrata
Između PSU	$N-1$	$SSB = \sum_{i=1}^{N}\sum_{j=1}^{M}(\bar{y}_{iU} - \bar{y}_U)^2$	MSB
Unutar PSU	$N(M-1)$	$SSW = \sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij} - \bar{y}_{iU})^2$	MSW
Ukupno	$NM-1$	$SSTO = \sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij} - \bar{y}_U)^2$	S^2

Tabela 1. Populaciona ANOVA tabela za klaster uzorkovanje

Kod klaster uzorkovanja je obrnuto. U jednofaznom klaster uzorkovanju promenljivost nepristrasnog ocenjivača od t u potpunosti zavisi od promenljivosti između klastera jer

$$S_t^2 = \sum_{i=1}^{N} \frac{(t_i - \bar{t}_U)^2}{N-1} = \sum_{i=1}^{N} \frac{M^2(\bar{y}_{iU} - \bar{y}_U)^2}{N-1} = M(MSB).$$

Prema tome, za klaster uzorkovanje važi

$$V(\hat{t}_{cluster}) = N^2(1 - \frac{n}{N})\frac{M(MSB)}{n}. \tag{4.1}$$

Ukoliko je kod klaster uzorkovanja količnik MSB/MSW velik, tada klaster uzorkovanje smanjuje preciznost. U toj situaciji MSB je relativno velika jer meri promenljivost između klastera. Elementi u različitim klasterima se često razlikuju više nego elementi u istom klasteru jer različiti klasteri imaju različite sredine. Na primer, ako bismo uzeli klaster uzorak razreda i ispitivali

sve učenike izabranih razreda, primetili bismo da se prosečne fizičke sposobnosti menjaju od razreda do razreda. Izuzetan nastavnik fizičkog vaspitanja može poboljšati fizičke sposobnosti čitavog razreda. Sa druge strane, učenici razreda iz siromašnog kraja mogu biti nedovoljno uhranjeni i imati slabije razvijene fizičke sposobnosti. Nemerljivi faktori, poput veštine predavanja ili siromaštva, mogu uticati na srednju vrednost klastera i na taj način prouzrokovati veliku MSB.

Unutar razreda, takođe, učenici poseduju različite fizičke sposobnosti. MSW je zbirna vrednost disperzija unutar klastera. U slučaju da su klasteri relativno homogeni, na primer učenici istog razreda imaju slične fizičke sposobnosti, MSW će biti mala.

Uporedimo sada klaster uzorkovanje sa prostim slučajnim uzorkovanjem. Ako bismo umesto klaster uzorka od M elemenata iz n klastera uzeli prost slučajan uzorak veličine nM, disperzija ocenjivača ukupne vrednosti bi bila

$$V(\hat{t}_{SRS}) = (NM)^2(1 - \frac{nM}{NM})\frac{S^2}{nM} = N^2(1 - \frac{n}{N})\frac{MS^2}{n}.$$

Upoređujući ovu formulu sa (4.1) zaključujemo da je klaster uzorkovanje manje efikasno od prostog slučajnog uzorkovanja ukoliko je $MSB > S^2$.

Unutarklasni koeficijent korelacije (eng. Intraclass Correlation Coefficient - ICC) pokazuje koliko su slični elementi unutar istog klastera. On predstavlja meru homogenosti unutar klastera. ICC se definiše kao Pearsonov koeficijent korelacije za $NM(M-1)$ parova (y_{ij}, y_{ik}), $i = 1, 2, ..., N$, $j \neq k$, i na osnovu veličina iz ANOVA tebele populacije, može se računati kao

$$ICC = 1 - \frac{M}{M-1}\frac{SSW}{SSTO}. \tag{4.2}$$

Pošto je $0 \leq SSW/SSTO \leq 1$, sledi

$$-\frac{1}{M-1} \leq ICC \leq 1.$$

Ukoliko su klasteri savršeno homogeni, tada je $SSW = 0$, pa je $ICC = 1$. Relacija (4.2) takođe implicira

$$MSB = \frac{NM-1}{M(N-1)}S^2[1 + (M-1)ICC].$$

Postavlja se pitanje koliko gubimo na preciznosti koristeći klaster uzorak. Iz prethodnih jednačina znamo da važi

$$\frac{V(\hat{t}_{cluster})}{V(\hat{t}_{SRS})} = \frac{MSB}{S^2} = \frac{NM-1}{M(N-1)}[1 + (M-1)ICC]. \tag{4.3}$$

Ako je broj PSU u populaciji - N veliko tako da $NM - 1 \approx M(N-1)$, tada je odnos disperzija iz (4.3) približno jednak $1 + (M-1)ICC$. Prema tome, $1 + (M-1)ICC$ sekundarnih uzoračkih jedinica uzetih u jednofaznom klaster uzorku daje približno istu količinu informacija kao jedna SSU uzeta prostim slučajnim uzorkovanjem. Ako je $ICC = \frac{1}{2}$ i $M = 5$, tada je $1 + (M-1)ICC = 3$, pa bi trebalo ispitati 300 elemenata klaster uzorka da bi se dobila ista preciznost kao kod 100 elemenata prostog slučajnog uzorka. Ipak, pošto je često znatno jeftinije i jednostavnije prikupljati podatke iz klaster uzorka, ostaje nada da klaster uzorkovanje pruža veću preciznost po potrošenoj novčanoj jedinici.

Već je napomenuto da ICC predstavlja meru homogenosti klastera. ICC je pozitivan ukoliko su elementi unutar PSU slični. Tada je SSW malo u poređenju sa SSTO, pa je ICC relativno velik. Kada je ICC pozitivan, klaster uzorkovanje je manje efikasno od prostog slučajnog uzorkovanja.

Ukoliko se klasteri prirodno javljaju u populaciji, ICC je uglavnom pozitivan. Elementi istog klastera su obično sličniji od slučajno izabranih elemenata iz populacije jer se elementi istog klastera nalaze u sličnoj sredini. Na primer, za očekivati je da lekovita vrela, koja se nalaze u istom geografskom klasteru, imaju sličan nivo pesticida, a da jedan deo grada ima različitu učestalost malih boginja od drugog dela grada. U ljudskoj populaciji, lični izbor kao i interakcija između članova domaćinstva ili komšija može uzrokovati da ICC bude pozitivan. Bogata domaćinstva teže da žive u njima materijalno sličnim okruženjima, a osobe iz iste sredine mogu deliti i neka slična mišljenja.

ICC je negativan ukoliko se elementi unutar klastera više razlikuju nego slučajno izabrani elementi iz cele populacije. Ova pojava uzrokuje da sredine klastera budu približno jednake, jer $SSTO = SSW + SSB$, pa ukoliko je SSTO fiksirana vrednost, a SSW velika, onda SSB mora biti mala. Ukoliko je $ICC < 0$, klaster uzorkovanje je efikasnije od prostog slučajnog uzorkovanja elemenata. ICC je retko negativan u klasterima koji se javljaju prirodno. Negativne vrednosti se mogu javiti u nekim sistematskim uzorcima ili veštačkim klasterima o kojima će kasnije biti reči.

ICC je definisan samo za klastere jednakih veličina. Analogna veličina koja se može koristiti kao mera homogenosti populacije u opštem slučaju je prilagođeno R^2, u oznaci R_a^2 koje je definisano kao

$$R_a^2 = 1 - \frac{MSW}{S^2}.$$

Ukoliko su svi klasteri jednakih veličina, povećanje disperzije usled klaster uzorkovanja je

$$\frac{MSB}{S^2} = 1 + \frac{N(M-1)}{N-1} R_a^2.$$

Poređenjem ove formule sa (4.3), može se uočiti da je za većinu populacija R_a^2 bliska vrednosti ICC. R_a^2 je sa razlogom mera homogenosti zbog svoje interpretacije u linearnoj regresiji: to je relativna promenljivost u populaciji objašnjena sredinama klastera i prilagođena broju stepeni slobode. Ako su klasteri homogeni, onda se sredine klastera značajno razlikuju u poređenju sa razlikama unutar klastera i u tom slučaju je R_a^2 veliko.

Primer 4.3 *Veštačke populacije*

Razmotrimo dve veštačke populacije od kojih svaka ima po tri klastera sa po tri elementa.

	Populacija A			*Populacija B*		
Klaster 1	*10*	*20*	*30*	*9*	*10*	*11*
Klaster 2	*11*	*20*	*32*	*17*	*20*	*20*
Klaster 3	*9*	*17*	*31*	*31*	*32*	*30*

Elementi ove dve populacije su isti, tako da obe populacije imaju istu srednju vrednost i disperziju

$$\bar{y}_U = 20, \qquad S^2 = 84.5.$$

U populaciji A veća varijabilnost se javlja unutar klastera nego između klastera, dok je kod populacije B obrnuto, veća različitost se uočava između klastera nego unutar klastera.

	Populacija A		*Populacija B*	
	$\bar{y}_{iU}$	S_i^2	$\bar{y}_{iU}$	S_i^2
Klaster 1	*20*	*100*	*10*	*1*
Klaster 2	*21*	*111*	*19*	*3*
Klaster 3	*19*	*124*	*31*	*1*

ANOVA tabela za populaciju A:

Izvor	*df*	*SS*	*MS*	*F*
Između PSU	*2*	*6*	*3*	*0.03*
Unutar PSU	*6*	*670*	*111.7*	
Ukupno	*8*	*676*	*84.5*	

ANOVA tabela za populaciju B:

Izvor	*df*	*SS*	*MS*	*F*
Između PSU	*2*	*666*	*333*	*199.8*
Unutar PSU	*6*	*10*	*1.7*	
Ukupno	*8*	*676*	*84.5*	

Za populaciju A je

$$R_a^2 = -0.3215$$

i

$$ICC = 1 - (\frac{3}{2})\frac{670}{676} = -0.4867.$$

U populaciji A uočavamo velike razlike između elemenata unutar klastera, a male razlike između srednjih vrednosti klastera. Iz tog razloga, vrednosti ICC i R_a^2 su negativne. Elementi istog klastera su manje slični od slučajno izabranih elemenata iz cele populacije. Zbog toga je klaster uzorkovanje u ovom slučaju efikasnije od prostog slučajnog uzorkovanja.

Za populaciju B je

$$R_a^2 = 0.9803$$

i

$$ICC = 1 - (\frac{3}{2})\frac{10}{676} = 0.9778.$$

Kod populacije B je obrnuto. Velike razlike uočavamo između klastera, dok elementi unutar klastera imaju približno jednake vrednosti. Vrednosti ICC i R_a^2 su približno jednake 1. Ovo je primer populacije za koju je klaster uzorkovanje manje efikasno od prostog slučajnog uzorkovanja. □

Većina populacija, u stvarnom životu, se nalazi između ova dva ekstrema. Kod njih je ICC najčešće pozitivna vrednost, ali ne bliska 1. Dakle, klaster uzorkovanje, u opštem slučaju, nije najefikasniji metod uzorkovanja, ali se ta smanjena efikasnost nadoknađuje uštedom troškova.

4.2.3 Klasteri različitih veličina

U društvenim istraživanjima klasteri su retko jednakih veličina. Ukoliko bismo, na primer, želeli da ispitamo stopu nezaposlenosti u Novom Sadu deleći grad na blokove, a zatim ispitujući sve stanovnike slučajno izabranih blokova, primetili bismo da u različitim blokovima živi različit broj stanovnika, a samim tim postoje i velike razlike u veličinama blokova koji u ovom slučaju predstavljaju klastere.

U jednofaznom klaster uzorku obima n od N PSU, znamo kako da ocenimo ukupnu vrednost i srednju vrednost populacije na dva načina: korišćenjem nepristrasnih ocenjivača i korišćenjem ocenjivača količnika.

Nepristrasna ocena

Nepristrasan ocenjivač za ukupnu vrednost populacije t, kao i standardna greška, računaju se na isti način kao i kod klastera jednakih veličina, pa je

$$\hat{t}_{unb} = \frac{N}{n} \sum_{i \in \mathcal{S}} t_i$$

i

$$SE(\hat{t}_{unb}) = N \sqrt{(1 - \frac{n}{N}) \frac{s_t^2}{n}}.$$

Razlika između klastera jednakih i različitih veličina je u tome što je disperzija pojedinačnih ukupnih vrednosti populacije klastera t_i, u opštem slučaju velika kada klasteri imaju različite veličine. U primeru ispitivanja stope nezaposlenosti u Novom Sadu, zanima nas ukupan broj nezaposlenih osoba u gradu. Tada t_i predstavlja broj nezaposlenih osoba u bloku i. Za očekivati je da u više naseljenim blokovima postoji veći broj nezaposlenih osoba nego u manje naseljenim blokovima. Prema tome, verovatno će t_i biti veliko kada je veličina klastera M_i veća, a malo kada je M_i mala. Tada je često s_t^2 veće u klaster uzorku u kome su PSU različitih veličina, nego u slučaju kada PSU imaju isti broj SSU.

Verovatnoća da se određena PSU nalazi u uzorku je n/N, s obzirom da se uzima prost slučajan uzorak obima n od N PSU. Pošto se primenjuje jednofazno klaster uzorkovanje, određena SSU će biti uključena u uzorak samo kada je u uzorak uključena PSU kojoj ona pripada. Prema tome, težinski koeficijent je

$$w_{ij} = \frac{1}{P\{j\text{-ta SSU iz } i\text{-te PSU je u uzorku}\}} = \frac{N}{n}.$$

Jednofazno klaster uzorkovanje daje samotežinski uzorak kada se PSU biraju sa jednakim verovatnoćama. Koristeći težinske koeficijente, ocenjivač ukupne vrednosti populacije može biti zapisan na sledeći način

$$\hat{t}_{unb} = \sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{S}_i} w_{ij} y_{ij}.$$

Formule za $\hat{t}_{unb}$ i $SE(\hat{t}_{unb})$ možemo koristiti za izvođenje nepristrasnog ocenjivača za srednju vrednost populacije $\bar{y}_U$ i njegovu disperziju. Neka je

$$K = \sum_{i=1}^{N} M_i$$

ukupan broj SSU u populaciji. Tada je

$$\hat{\bar{y}}_{unb} = \frac{\hat{t}_{unb}}{K},$$

$$SE(\hat{\bar{y}}_{unb}) = \frac{SE(\hat{t}_{unb})}{K}.$$

Međutim, za upotrebu ovih formula neophodno je znati K, a vrlo često je M_i poznato samo za one klastere koji su u uzorku, a ne i za sve ostale, što znači da je K nepoznato.

Ocena količnika

Obično se očekuje da je ukupna vrednost populacije i-tog klastera t_i u korelaciji sa veličinom tog klastera M_i, pa se vrednosti M_i koriste kao pomoćne promenljive. Tada se srednja i ukupna vrednost populacije mogu oceniti pomoću

$$\hat{\bar{y}}_r = \frac{\sum_{i\in\mathcal{S}} t_i}{\sum_{i\in\mathcal{S}} M_i}$$

i

$$\hat{t}_r = K\hat{\bar{y}}_r.$$

Imenilac zavisi od toga koje konkretne PSU su uključene u uzorak, tako da i brojilac i imenilac variraju od uzorka do uzorka. Iz formule (3.3) proizilaze standardne greške

$$SE(\hat{\bar{y}}_r) = \sqrt{(1-\frac{n}{N})\frac{1}{n\bar{M}_U^2}\frac{\sum_{i\in\mathcal{S}}(t_i - \hat{\bar{y}}_r M_i)^2}{n-1}} \tag{4.4}$$

$$= \sqrt{(1-\frac{n}{N})\frac{1}{n\bar{M}_U^2}\frac{\sum_{i\in\mathcal{S}} M_i^2(\bar{y}_i - \hat{\bar{y}}_r)^2)}{n-1}}$$

i

$$SE(\hat{t}_r) = N\sqrt{(1-\frac{n}{N})\frac{1}{n}\frac{\sum_{i\in\mathcal{S}} M_i^2(\bar{y}_i - \hat{\bar{y}}_r)^2}{n-1}}.$$

Ukoliko je prosečna veličina klastera u populaciji $\bar{M}_U = K/N$ nepoznata, ona se može zameniti sa prosečnom veličinom klastera u uzorku. Drugim rečima, $\bar{M}_U$ se može zameniti sa $\bar{M}_{\mathcal{S}}$ u formuli (4.4).

Ocena $\hat{\bar{y}}_r$ se može računati i pomoću težina w_{ij} kao

$$\hat{\bar{y}}_r = \frac{\sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}_i} w_{ij}y_{ij}}{\sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}_i} w_{ij}}.$$

Disperzija ocenjivača količnika zavisi od promenljivosti sredina po elementu u klasterima i može biti dosta manja od disperzije nepristrasnog ocenjivača, što znači da su često ocene dobijene na osnovu količnika preciznije od ocena dobijenih na osnovu nepristrasnih ocenjivača. Međutim, treba napomenuti da je za $\hat{t}_r$ potrebno znati ukupan broj elemenata populacije K, dok za nepristrasni ocenjivač to nije neophodno.

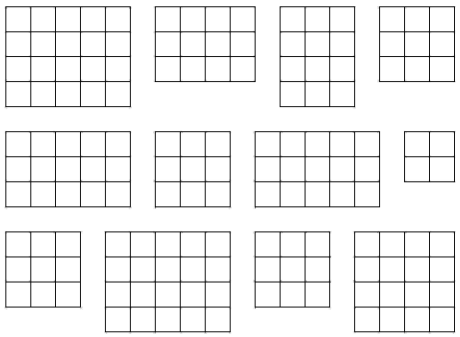
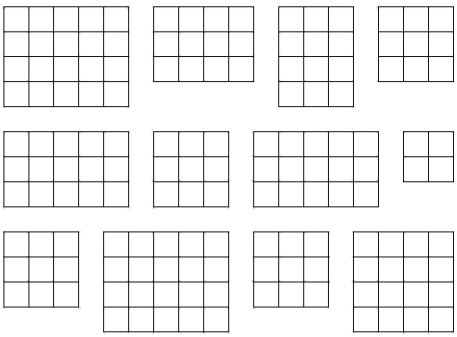
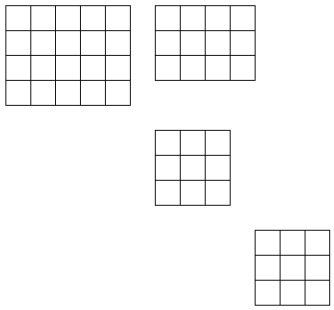
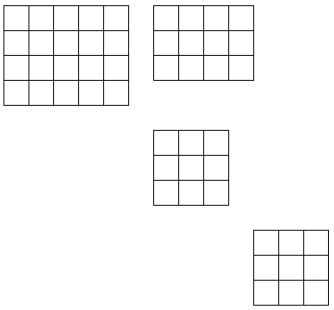
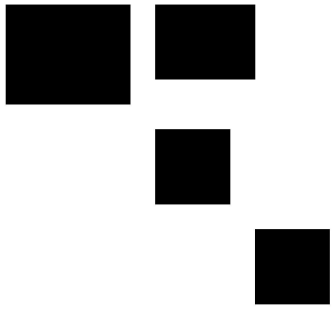
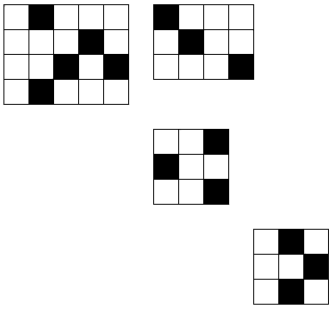
4.3 Dvofazno klaster uzorkovanje

U jednofaznom klaster uzorkovanju ispituju se sve SSU unutar izabranih PSU (klastera). Ipak, u mnogim situacijama elementi u klasteru mogu biti veoma slični pa ispitivanje svih sekundarnih uzoračkih jedinica u PSU može biti skupo i nepotrebno. Osim toga, merenje SSU može biti skupo u poređenju sa cenom uzorkovanja PSU. U ovim situacijama, uzimanje poduzorka iz svake izabrane PSU može biti daleko jeftinije pa se tada primenjuje dvofazno klaster uzorkovanje. Klasteri se biraju slučajno, a zatim se iz svakog izabranog klastera slučajno biraju elementi. Ovaj tip uzorkovanja čine dve faze:

1. Bira se SRS $\mathcal{S}$ od n PSU iz populacije od N PSU.

2. Bira se SRS SSU iz svake izabrane PSU. SRS od m_i elemenata iz i-tog klastera se označava sa $\mathcal{S}_i$.

Razlike između jednofaznog i dvofaznog klaster uzorkovanja su prikazane na sledećoj slici.

Jednofazno klaster uzorkovanje	Dvofazno klaster uzorkovanje
Populacija od N PSU: 	Populacija od N PSU: 
Uzima se SRS od n PSU: 	Uzima se SRS od n PSU: 
Uzorkuju se sve SSU u izabranim PSU: 	Uzima se SRS od m_i SSU u i-toj izabranoj PSU: 

U dvofaznom klaster uzorku notacija i ocenjivači postaju još složeniji nego u jednofaznom klaster uzorku, jer treba uzeti u obzir različitosti koje dolaze iz obe faze prikupljanja podataka. Ocenjivanje t i $\bar{y}_U$ je analogno onom u jednofaznom klaster uzorku, ali formule za disperziju postaju znatno komplikovanije.

Nepristrasna ocena

Kod jednofaznog klaster uzorka nepristrasni ocenjivač ukupne vrednosti populacije je bio oblika

$$\hat{t}_{unb} = \frac{N}{n} \sum_{i \in \mathcal{S}} t_i,$$

pri čemu su t_i bile poznate vrednosti jer smo uzorkovali sve SSU u izabranoj PSU. U dvofaznom klaster uzorku, međutim, s obzirom da ne uzimamo sve SSU iz izabrane PSU, moramo oceniti individualne ukupne vrednosti populacije PSU na sledeći način

$$\hat{t}_i = \sum_{j \in \mathcal{S}_i} \frac{M_i}{m_i} y_{ij} = M_i \bar{y}_i.$$

Sada je nepristrasni ocenjivač ukupne vrednosti populacije

$$\hat{t}_{unb} = \frac{N}{n} \sum_{i \in \mathcal{S}_i} \hat{t}_i = \frac{N}{n} \sum_{i \in \mathcal{S}} M_i \bar{y}_i. \tag{4.5}$$

Pri dvofaznom klaster uzorkovanju $\hat{t}_i$ su slučajne promenljive. Zbog toga, disperzija ocenjivača ukupne vrednosti populacije ima dve komponente:

1. promenljivost između PSU
2. promenljivost između SSU unutar PSU.

U jednofaznom klaster uzorkovanju ne moramo da brinemo o drugoj komponenti.

Disperzija od $\hat{t}_{unb}$ je jednaka disperziji od $\hat{t}_{unb}$ iz jednofaznog klaster uzorka uvećanoj za dodatni član, s obzirom da $\hat{t}_i$ ocenjuju ukupne vrednosti populacija klastera. Za dvofazno klaster uzorkovanje važi

$$V(\hat{t}_{unb}) = N^2 (1 - \frac{n}{N}) \frac{S_t^2}{n} + \frac{N}{n} \sum_{i=1}^{N} (1 - \frac{m_i}{M_i}) M_i^2 \frac{S_i^2}{m_i},$$

gde je

- S_t^2 populaciona disperzija ukupnih vrednosti klastera
- S_i^2 populaciona disperzija elemenata unutar i-tog klastera.

Prvi sabirak u prethodnoj formuli predstavlja disperziju iz jednofaznog klaster uzorkovanja, dok je drugi sabirak dodatna disperzija usled poduzorkovanja.

Za ocenjivanje $V(\hat{t}_{unb})$ koristimo uzoračke disperzije

$$s_t^2 = \frac{\sum_{i\in\mathcal{S}}(\hat{t}_i - \frac{\hat{t}_{unb}}{N})^2}{n-1},$$

$$s_i^2 = \frac{\sum_{j\in\mathcal{S}_i}(y_{ij} - \bar{y}_i)^2}{m_i - 1}. \tag{4.6}$$

Može se pokazati da je nepristrasan ocenjivač disperzije oblika

$$\hat{V}(\hat{t}_{unb}) = N^2(1 - \frac{n}{N})\frac{s_t^2}{n} + \frac{N}{n}\sum_{i\in S}(1 - \frac{m_i}{M_i})M_i^2\frac{s_i^2}{m_i}.$$

Standardna greška $SE(\hat{t}_{unb})$, je naravno kvadratni koren navedene ocenjene disperzije. U mnogim slučajevima N/n je malo u poređenju sa N^2, pa je doprinos drugog sabirka ocenjivaču disperzije neuporedivo manji od doprinosa prvog sabirka.

Ukoliko znamo ukupan broj elemenata u populaciji K, možemo oceniti srednju vrednost populacije sa

$$\hat{\bar{y}}_{unb} = \frac{\hat{t}_{unb}}{K}, \tag{4.7}$$

pri čemu je standardna greška

$$SE(\hat{\bar{y}}_{unb}) = \frac{SE(\hat{t}_{unb})}{K}.$$

Kao i u jednofaznom klaster uzorku sa različitim veličinama, komponenta disperzije koja potiče od varijabilnosti između PSU može biti veoma velika pošto na nju utiču i razlike u veličini M_i i razlike u $\bar{y}_i$. Ako su veličine klastera različite, ova komponenta je velika, čak i kada su sredine klastera približno jednake. Pored nepristrasnih ocenjivača u dvofaznom klaster uzorkovanju mogu se koristiti i ocenjivači količnika.

Ocena količnika

Srednja vrednost populacije se može oceniti korišćenjem ocenjivača količnika. Na osnovu oznaka iz trećeg poglavlja, za vrednosti y_i uzimaju se ocenjene ukupne vrednosti populacije klastera $\hat{t}_i$, dok su vrednosti x_i veličine klastera M_i, pa je

$$\hat{\bar{y}}_r = \frac{\sum_{i\in\mathcal{S}} \hat{t}_i}{\sum_{i\in\mathcal{S}} M_i} = \frac{\sum_{i\in\mathcal{S}} M_i\bar{y}_i}{\sum_{i\in\mathcal{S}} M_i}. \tag{4.8}$$

Može se pokazati da je disperzija oblika

$$\hat{V}(\hat{\bar{y}}_r) = \frac{1}{\bar{M}^2}[(1-\frac{n}{N})\frac{s_r^2}{n} + \frac{1}{nN}\sum_{i\in\mathcal{S}} M_i^2(1-\frac{m_i}{M_i})\frac{s_i^2}{m_i}],$$

gde je s_i^2 definisano sa (4.6), a

$$s_r^2 = \frac{\sum_{i\in\mathcal{S}}(M_i\bar{y}_i - M_i\hat{\bar{y}}_r)^2}{n-1},$$

dok je $\bar{M}$ prosečna veličina klastera.

Ukoliko su svi M_i jednaki, nepristrasni ocenjivač je zapravo jednak ocenjivaču količnika. Ukoliko se veličine klastera M_i razlikuju, nepristrasni ocenjivač često ne funkcioniše najbolje i tada je bolje koristiti ocenjivač količnika. Naredni primer ilustruje da nepristrasni ocenjivač $\hat{t}_{unb}$ može imati veliku disperziju kada se veličine klastera veoma razlikuju.

Primer 4.4 *Slučaj šestonogih kučića*

Pretpostavimo da želimo da ocenimo prosečan broj nogu zdravih kučića iz prihvatilišta u jednom gradu. U gradu postoje dva prihvatilišta za kučiće: prihvatilište A, sa 30 kučića, i prihvatilište B, sa 10 kučića. Neka je verovatnoća izbora jednaka za oba prihvatilišta i iznosi 1/2. Nakon odabira prihvatilišta, treba iz njega slučajnim odabirom izabrati 2 kučeta i, koristeći $\hat{\bar{y}}_{unb}$, oceniti prosečan broj nogu po kučetu.

Pretpostavimo da je slučajno izabrano prihvatilište A. Naravno, svaki od 2 izabrana psa ima 4 noge, pa je $\hat{t}_A = 30 \times 4 = 120$. Na osnovu (4.5) i (4.7), nepristrasna ocena za ukupan broj nogu svih kučića u oba doma je

$$\hat{t}_{unb} = \frac{2}{1}\hat{t}_A = 240.$$

Ako podelimo ocenjen ukupan broj nogu sa brojem kučića da bismo ocenili prosečan broj nogu po kučetu, dobijamo $240/40 = 6$.

Ako je pak slučajno izabrano prihvatilište B, tada je $\hat{t}_B = 10 \times 4 = 40$ *i nepristrasna ocena ukupnog broja nogu svih kučića je*

$$\hat{t}_{unb} = \frac{2}{1}\hat{t}_B = 80,$$

a nepristrasna ocena prosečnog broja nogu po kučetu je $80/40 = 2$. *Zaključujemo da ovo nisu dobre ocene za prosečan broj nogu po kučetu. Međutim, ocenjivač je matematički nepristrasan jer je* $(6 + 2)/2 = 4$, *tako da prosečna vrednost svih mogućih uzoraka rezultuje tačnom vrednošću 4. Slabost ovog ocenjivača se odražava u veoma velikoj disperziji*

$$V(\hat{t}_{unb}) = (1 - \frac{1}{2})2^2\frac{S_t^2}{1} + \frac{2}{1}\sum_{i=1}^{2}(1 - \frac{m_i}{M_i})M_i^2\frac{S_i^2}{m_i} = \frac{1}{2}(4)(3200) = 6400.$$

U ovom primeru, ocenjivač količnika je mnogo bolji. U slučaju odabira prihvatilišta A dobija se $\hat{\bar{y}}_r = 120/30 = 4$, *a ukoliko je prihvatilište B izabrano,* $\hat{\bar{y}}_r = 40/10 = 4$. *Pošto je ocena ista za sve moguće uzorke, sledi da je disperzija* $V(\hat{\bar{y}}_r) = 0$ *pa je u ovom primeru ocena količnika preciznija od nepristrasne ocene.* □

U opštem slučaju, nepristrasan ocenjivač ukupne vrednosti populacije nije efikasan ako su veličine klastera M_i različite i ako je t_i grubo proporcionalno sa M_i. Disperzija od $\hat{t}_{unb}$ zavisi od disperzije od t_i i ona može biti velika ako su M_i različiti.

Za razliku od nepristrasnog ocenjivača, ocenjivač količnika u opštem slučaju funkcioniše dobro kada su t_i grubo proporcionalni sa M_i. Podsetimo se, iz ocene količnika, MSE ocenjivača $\hat{B}$ je proporcionalna disperziji reziduala iz modela. Kada se koristi ocena količnika u klaster uzorkovanju, MSE ocenjivača $\hat{\bar{y}}_r$ je proporcionalna sa $\sum_{i=1}^{N}(t_i - \bar{y}_U M_i)^2$. Kada su glavne promenljive t_i u jakoj pozitivnoj korelaciji sa pomoćnim promeljivama M_i, reziduali su mali. U primeru sa šestonogim kučićima, ukupan broj nogu u prihvatilištu t_i, je tačno četiri puta veći od ukupnog broja kučića u prihvatilištu M_i, tako da je disperzija ocenjivača količnika jednaka nuli i zato je u tom primeru mnogo efikasnije koristiti ocenu količnika.

Ovo je veoma važno napomenuti, s obzirom da su mnogi prirodni klasteri obično različitih veličina, pa je najčešće ukupna vrednost populacije klastera proporcionalna broju SSU u njemu. Na primer, u klaster uzorku bolnica, očekujemo da će u bolnici sa 500 bolesnika biti veći broj zadovoljnih bolesnika nego u bolnici sa 20 bolesnika, iako procenat zadovoljnih bolesnika u obe bolnice može biti isti. Na primer, ukupna vrednost ocena iz matematike za

sve učenike razreda biće znatno veća za razrede sa većim brojem učenika nego za malobrojnije razrede. U opštem slučaju, očekujemo da pronađemo više pčela u većoj sredini nego u manjoj. U svim ovim situacijama, dok ocenjivač $\hat{\bar{y}}_r$ funkcioniše dobro, ocenjivač $\hat{t}_{unb}$ obično ima veliku disperziju.

4.4 Korišćenje težina u klaster uzorcima

Većina statističara koristi uzoračke težine za ocenjivanje ukupne i srednje vrednosti populacije u klaster uzorcima. Pomoću težina mogu se naći ocene gotovo svih značajnih veličina iz proizvoljnog verovatnosnog uzorka. Iz tog razloga, one predstavljaju značajan alat za analizu podataka istraživanja.

Težina elementa je recipročna vrednost verovatnoće izbora elementa. Kod klaster uzorka, verovatnoća izbora SSU je

$$P(\text{izabrana } j\text{-ta SSU iz } i\text{-te PSU})$$

$$= P(\text{izabrana } i\text{-ta PSU}) \times P(\text{izabrana } j\text{-ta SSU} \mid \text{izabrana } i\text{-ta PSU})$$

$$= \frac{n}{N}\frac{m_i}{M_i}.$$

Prema tome, težina je

$$w_{ij} = \frac{NM_i}{nm_i}.$$

Težina w_{ij} određuje broj elemenata populacije koje predstavlja j-ta jedinica iz i-tog klastera u uzorku. Ako su, na primer, PSU blokovi, a SSU domaćinstva, onda domaćinstvo j u bloku i predstavlja $\frac{NM_i}{nm_i}$ domaćinstava populacije i to sebe i još $(\frac{NM_i}{nm_i} - 1)$ drugih domaćinstava. Tada je

$$\hat{t}_{unb} = \sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}_i} w_{ij}y_{ij}$$

i

$$\hat{\bar{y}}_r = \frac{\hat{t}_{unb}}{\sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}_i} w_{ij}}.$$

Uočava se da je $\hat{t}_{unb}$ isto kao u (4.5), a $\hat{\bar{y}}_r$ isto kao u (4.8). Uzoračke težine samo obezbeđuju zgodan način za računanje ovih ocena, ali ne eliminišu nedostatke poput velikih disperzija. Osim toga, uzoračke težine ne pružaju informaciju za pronalaženje standardnih grešaka, za čije računanje se moraju koristiti ili navedene formule ili neke druge metode.

U dvofaznom klaster uzorkovanju, kod samotežinskog dizajna, svaka SSU predstavlja isti broj SSU u populaciji. Na primer, za samotežinski uzorak stanovnika Beograda možemo prvo uzeti prost slučajan uzorak opština Beograda, a onda uzeti uzorak obima m_i od ukupno M_i stanovnika opštine i. Kako bi svaki stanovnik u uzorku predstavljao isti broj stanovnika populacije, m_i mora biti proporcionalno sa M_i, tako da m_i/M_i treba biti približno konstantno. Prema tome, iz naseljenijih opština treba uzorkovati više stanovnika nego iz manje naseljenih opština.

4.5 Dizajniranje klaster uzorka

Organizacije i agencije koje sprovode skupa istraživanja velikih razmera moraju posvetiti dosta vremena dizajniranju istraživanja. Za dizajniranje i testiranje nekih istraživanja potrebno je i po nekoliko godina. Čak i onda važi osnovni princip dizajniranja istraživanja koji glasi: "Istraživanje se može najbolje dizajnirati nakon što je završeno". Tek nakon završenog istraživanja može se proceniti efekat korišćene određene metode uzorkovanja i tek tada se može saznati gde je trebalo više sredstava uložiti u cilju dobijanja boljih informacija i rezultata.

Što više informacija posedujemo o populaciji, to bolje možemo dizajnirati efikasno uzorkovanje. Ako znamo vrednosti y_{ij} za svaku jedinicu u našoj populaciji, možemo bez greške dizajnirati istraživanje za izučavanje populacije. Naravno, ono bi bilo nepotrebno, jer već sve znamo o populaciji. Međutim, ukoliko nam je malo informacija o populaciji poznato, verovatno ćemo nakon istraživanja prikupiti određene željene informacije, ali je takođe, moguće da nećemo imati najefikasniji dizajn za naše istraživanje. Međutim, naša novostečena saznanja o populaciji možemo iskoristiti za pravljenje efikasnijih istraživanja u budućnosti.

Prilikom dizajniranja klaster uzorka, potrebno je voditi računa o sledećim pitanjima:

1. Kolika ukupna preciznost je potrebna?
2. Koje veličine treba da budu PSU?
3. Koliko SSU treba uzorkovati iz svake PSU koja je izabrana u uzorak?
4. Koliko PSU treba uzorkovati?

Sa prvim pitanjem se moramo suočiti u svakom istraživanju. Da bismo odgovorili na pitanja 2, 3 i 4 potrebno je da znamo troškove uzorkovanja

PSU za moguće veličine PSU, troškove uzorkovanja SSU i meru homogenosti (R_a^2 ili ICC) za moguće veličine PSU.

Određivanje veličine PSU

Veličina PSU tj. klastera je najčešće prirodna stvar. Razredi u školi, ili bolnice sa bolesnicima su bili primeri takvih prirodnih klastera.

Međutim, postoje istraživanja u kojima istraživač ima širok izbor za određivanje veličine PSU. Na primer, u istraživanju procene pola i starosti jelena u regijama Kolorada, PSU mogu biti određena područja, dok SSU mogu biti individualni jeleni ili grupe jelena u tim područjima. Ali, takođe treba odrediti veličinu tih područja, tj. treba odrediti da li veličina PSU treba biti $1km^2$, $2km^2$ ili $100km^2$.

Opšti princip u takvim istraživanjima je da što su veći klasteri, veća je i varijabilnost unutar njih, pa se tad povećava efikasnost klaster uzorkovanja. Međutim, ukoliko su klasteri isuviše veliki, tada se gubi ušteda koja je tipična za klaster uzorkovanje i koja predstavlja jednu od osnovnih prednosti ovog tipa uzorkovanja.

Bellhouse (1984) je dao pregled optimalnog dizajniranja uzorka, a navedena teorija pruža koristan vodič za dizajniranje sopstvenog istraživanja. Postoje mnogi načini isprobavanja različitih veličina PSU pre sprovođenja samog istraživanja. Jedan od njih je modeliranje veze između R_a^2 ili MSW i M i prilagođavanje modela uz pomoć preliminarnih podataka ili informacija iz drugih istraživanja, a nakon toga upoređivanje troškova za različite kombinacije R_a^2 i M. Drugi način je vršenje eksperimenta i prikupljanje podataka o relativnim troškovima i disperzijama za različite veličine PSU.

Primer 4.5 *Procenjivanje broja krompirovih zlatica*

Proučavane su različite veličine uzoračkih jedinica radi procene broja krompirovih zlatica. Uzorkovano je 10 slučajno odabranih lokacija sa svakog od 10 polja. Istraživači su vizuelno ispitivali svaku lokaciju i beležili male larve, velike larve i odrasle jedinke na svim listovima jednog stabla na svakoj od 5 susednih biljaka. Zatim su posmatrali različite veličine PSU - od jednog do pet stabala po lokaciji. Za dizajn "1 stablo/lokaciji" prikupljani su podaci sa jednog stabla po lokaciji, za dizajn "2 stabla/lokaciji" prikupljani su podaci sa dva stabla po lokaciji itd. Za sve šetnje među lokacijama, bilo im je potrebno 30min, a za ispitivanje jednog stabla 10sec. Prema tome, ukupno vreme za ispitivanje jednog stabla po lokaciji na svih 10 lokacija je bilo 30+100/60=31.67min. Pošto su troškovi uzorkovanja dodatnog stabla po lokaciji mali u odnosu na vreme potrebno za prelazak na drugo polje,

istraživači su došli do zaključka da je najefikasniji dizajn za procenu broja krompirovih zlatica dizajn "5 stabala/lokaciji".

Određivanje obima poduzorka (broja SSU u izabranoj PSU)

Osnovni cilj prilikom dizajniranja uzorka je dobijanje što više informacija uz što manje troškova. Ograničićemo se na dizajniranje dvofaznog klaster uzorka u slučaju kada svi klasteri imaju jednak broj SSU. Jedan pristup, za ovakve klastere jednakih veličina, je minimiziranje disperzije za fiksiran iznos troškova. Ako je $M_i = M$ i $m_i = m$ za sve PSU, tada se može pokazati da je

$$V(\hat{\bar{y}}_{unb}) = (1 - \frac{n}{N})\frac{MSB}{nM} + (1 - \frac{m}{M})\frac{MSW}{nm}, \tag{4.9}$$

gde su MSB i MSW, respektivno, sredine kvadrata između i unutar PSU iz populacione ANOVA tabele 1.

Ako je MSW $= 0$, tada je $R_a^2 = 1$, pa svi elementi unutar klastera imaju vrednosti sredine klastera. U tom slučaju dovoljno je uzeti $m = 1$, jer ispitivanje više od jednog elementa po klasteru samo povećava troškove i troši vreme bez povećanja preciznosti, jer ne donosi nove informacije. Za druge vrednosti R_a^2 optimalna raspodela zavisi od relativnih troškova uzorkovanja PSU i SSU.

Neka je data jednostavna funkcija ukupnih troškova

$$C = c_1 n + c_2 nm,$$

gde je c_1 trošak po PSU (bez troškova merenja SSU), a c_2 trošak merenja svake SSU. Vrednosti

$$n = \frac{C}{c_1 + c_2 m}$$

i

$$m = \sqrt{\frac{c_1 M(MSW)}{c_2(MSB - MSW)}} = \sqrt{\frac{c_1 M(N-1)}{c_2(NM-1)}(\frac{1}{R_a^2} - 1)}$$

minimizuju disperziju $V(\hat{\bar{y}}_{unb})$ za fiksne ukupne troškove C, pri čemu je naravno potrebno znati barem grubu ocenu R_a^2. Međutim, često neke druge vrednosti m daju slične rezultate. Crtanje projektovanih disperzija ocena može dati više informacija od samo jednog izračunavanja fiksnog rešenja. Osim toga, grafički pristup dozvoljava da izvršimo tzv. senzitivnu analizu na dizajnu: Šta ako su troškovi ili funkcija troškova malo drugačiji? Šta ako se vrednost R_a^2 malo promeni? Ovim pristupom se, takođe, mogu ispitivati različite funkcije troškova.

Iako smo posmatrali samo klastere jednakih veličina kada je $M_i = M$, ovaj pristup možemo iskoristiti i u slučaju kada su M_i različite vrednosti. Tada treba zameniti M sa $\bar{M}$, gde je $\bar{M}$ prosečna veličina klastera i odrediti prosečnu veličinu poduzorka $\bar{m}$. Jedan način je da uzmemo $\bar{m}$ jedinica iz svakog klastera koji je izabran u uzorak, a drugi je da rasporedimo jedinice tako da je

$$\frac{m_i}{M_i} = const.$$

Ukoliko M_i ne variraju previše, ova procedura daje prihvatljiv dizajn. Ako su vrednosti M_i značajno različite, a t_i su u korelaciji sa M_i, klaster uzorak sa jednakim verovatnoćama neće biti dovoljno efikasan.

Određivanje obima uzorka (broja PSU)

Nakon određivanja veličine PSU i odnosa PSU i SSU, potrebno je odrediti i broj PSU koje treba uzorkovati - n. Dizajniranje klaster uzorka je iterativni proces koji čine sledeći koraci:

1. Odrediti željenu preciznost
2. Odrediti veličinu PSU i obim poduzorka
3. Ograničiti željenu disperziju
4. Odrediti n da se postigne željena preciznost
5. Ponoviti postupak (dodajući pomoćne promenljive za ocenu količnika) sve dok troškovi istraživanja ne budu unutar našeg budžeta.

Ukoliko su klasteri jednakih veličina, ignorišući fpc, na osnovu (4.9) dobija se

$$V(\hat{\bar{y}}_{unb}) \leq \frac{1}{n}[\frac{MSB}{M} + (1 - \frac{m}{M})\frac{MSW}{m}] = \frac{1}{n}v.$$

Tada je 95%-tni interval poverenja

$$\hat{\bar{y}}_{unb} \pm 1.96\sqrt{\frac{1}{n}v}.$$

Prema tome, da bi se postigao željeni interval poverenja poluširine e, n se određuje na sledeći način

$$n = \frac{1.96^2 v}{e^2},$$

pri čemu ovaj pristup podrazumeva posedovanje nekog saznanja o v, na primer na osnovu preliminarnih istraživanja.

4.6 Sistematsko uzorkovanje

U ovoj glavi razmatrano je sistematsko uzorkovanje, koje predstavlja specijalan slučaj klaster uzorkovanja. Pretpostavimo da želimo da uzmemo uzorak veličine 3 iz populacije koja broji 12 elemenata:

$$1 \quad 2 \quad 3 \quad 4 \quad 5 \quad 6 \quad 7 \quad 8 \quad 9 \quad 10 \quad 11 \quad 12$$

Za uzimanje sistematskog uzorka treba izabrati slučajan broj između 1 i 4, pa počevši od njega uzimati svaki četvrti element. Dakle, ova populacija sadrži 4 PSU (klastera) i to su:

$$\{1, 5, 9\} \quad \{2, 6, 10\} \quad \{3, 7, 11\} \quad \{4, 8, 12\}.$$

Sada uzimamo prost slučajan uzorak klastera obima 1, tj. slučajno biramo jedan klaster i to je sistematski uzorak.

U populaciji od NM elemenata, postoji N mogućih izbora sistematskog uzorka, od kojih je svaki veličine M. Posmatramo srednju vrednost samo onog klastera koji čini naš sistematski uzorak

$$\bar{y}_i = \bar{y}_{iU} = \hat{\bar{y}}_{sys}.$$

Jednofazno klaster uzorkovanje sa klasterima jednakih veličina daje nepristrasne ocene, pa je zato $E[\hat{\bar{y}}_{sys}] = \bar{y}_U$, tj. ocenjivač $\hat{\bar{y}}_{sys}$ je nepristrasan. Za prost sistematski uzorak biramo $n = 1$ od N klastera, pa je disperzija ocenjivača $\hat{\bar{y}}_{sys}$ oblika

$$V(\hat{\bar{y}}_{sys}) = (1 - \frac{1}{N})\frac{S_t^2}{M^2} = (1 - \frac{1}{N})\frac{MSB}{M} \approx \frac{S^2}{M}[1 + (M-1)ICC].$$

Na osnovu poznatih oznaka za klaster uzorkovanje, M je veličina sistematskog uzorka. Ignorišući fpc, zaključujemo da je sistematski uzorak precizniji od SRS veličine M, ako je ICC negativno. Sistematski uzorak je mnogo precizniji od prostog slučajnog uzorka kada je disperzija unutar mogućih sistematskih uzoraka (klastera) veća od disperzije cele populacije. U tom slučaju su i srednje vrednosti klastera sličnije. Ukoliko je pak varijabilnost elemenata unutar sistematskog uzorka mala u poređenju sa onima u populaciji (odnosno, $ICC > 0$), tada elementi u uzorku daju slične informacije i za očekivati je da sistematski uzorak ima veću disperziju od SRS i tada je SRS precizniji od sistematskog uzorka.

Međutim, pošto je $n = 1$, ne možemo dobiti nepristrasnu ocenu od $V(\hat{\bar{y}}_{sys})$. Za ocenjivanje disperzije neophodno je imati dodatnu informaciju o strukturi populacije. Posmatrajmo sada tri različite strukture populacije.

1. Redosled elemenata populacije u uzoračkom okviru je slučajan.

   Tada će sistematsko uzorkovanje proizvesti uzorak koji se najverovatnije ponaša kao SRS. U mnogim situacijama redosled elemenata populacije nije povezan sa karakteristikom koja se ispituje. Primer ovakvog slučaja je lista osoba u uzoračkom okviru poređanih po azbučnom redu. Nema razloga da se veruje da će osobe u sistematskom uzorku biti više ili manje slične od osoba prostog slučajnog uzorka. Očekujemo da je $ICC \approx 0$. U ovoj situaciji, prosto slučajno i sistematsko uzorkovanje daju slične rezultate. Za ocenu $V(\hat{\bar{y}}_{sys})$ tada se koriste formule iz prostog slučajnog uzorkovanja.

2. Uzorački okvir je u rastućem ili opadajućem poretku.

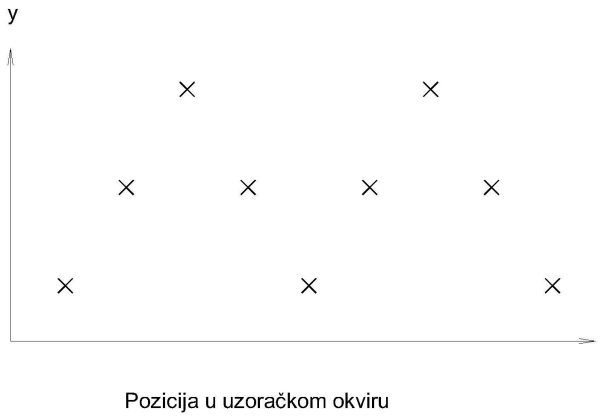
   U ovom slučaju sistematsko uzorkovanje je verovatno preciznije od prostog slučajnog uzorkovanja. Na primer, finansijska dokumentacija može biti poređana od najvećih iznosa na početku do najmanjih iznosa na kraju liste. Za takvu populaciju se kaže da ima pozitivnu autokorelaciju. Susedni elementi su sličniji od udaljenijih elemenata. U ovom slučaju, $V(\hat{\bar{y}}_{sys})$ je manja od disperzije uzoračke sredine prostog slučajnog uzorka iste veličine pošto je $ICC < 0$. Sistematski uzorak sadrži u ovom slučaju raširene elemente populacije, a moguće je da bi se SRS sastojao iz svih malih ili iz svih velikih vrednosti.

Kada je uzorački okvir u rastućem ili opadajućem poretku, formula za standardnu grešku prostog slučajnog uzorka se može koristiti, ali je moguće da će greška biti veća, a intervali poverenja konstruisani korišćenjem standardne greške prostog slučajnog uzorka preširoki.

Za populacije sa pozitivnom autokorelacijom stratifikovani uzorak je obično bolji od sistematskog. Ako je slučajno izabrani početni element previše blizu početku ili kraju uzoračkog intervala, sistematski uzorak će pružati ocenu koja je preniska ili previsoka.

3. Uzorački okvir je periodičan.

   U slučaju da uzorkovanje vršimo sa intervalom koji se poklapa sa periodičnošću, sistematsko uzorkovanje će biti manje precizno od prostog slučajnog uzorkovanja. Sistematsko uzorkovanje je najopasnije u slučaju kada je populacija u cikličnom ili periodičnom poretku, a uzorački interval se poklapa sa umnoškom tog perioda.



Na primer, pretpostavimo da su vrednosti populacije, redom:

$$1 \quad 2 \quad 3 \quad 1 \quad 2 \quad 3 \quad 1 \quad 2 \quad 3 \quad 1 \quad 2 \quad 3$$

a uzorački interval je 3. Tada će svi elementi u sistematskom uzorku biti jednaki. Ukoliko u ovom slučaju iskoristimo formulu SRS za ocenu disperzije, dobićemo $\hat{V}(\hat{\bar{y}}_{sys}) = 0$, a stvarna vrednost $\hat{V}(\hat{\bar{y}}_{sys})$ za ovu populaciju je 2/3. U ovom slučaju sistematski uzorak nije precizniji od jednog proizvoljnog posmatranja slučajno odabranog elementa iz populacije.

Sistematsko uzorkovanje se često koristi u slučajevima kada istraživač želi reprezentativan uzorak populacije, ali ne poseduje sredstva za konstruisanje uzoračkog okvira unapred. Obično se koristi za odabir elemenata u početnoj fazi klaster uzorkovanja. U mnogim situacijama u kojima se koristi sistematsko uzorkovanje, sistematski uzorak se može tretirati kao prost slučajan uzorak.

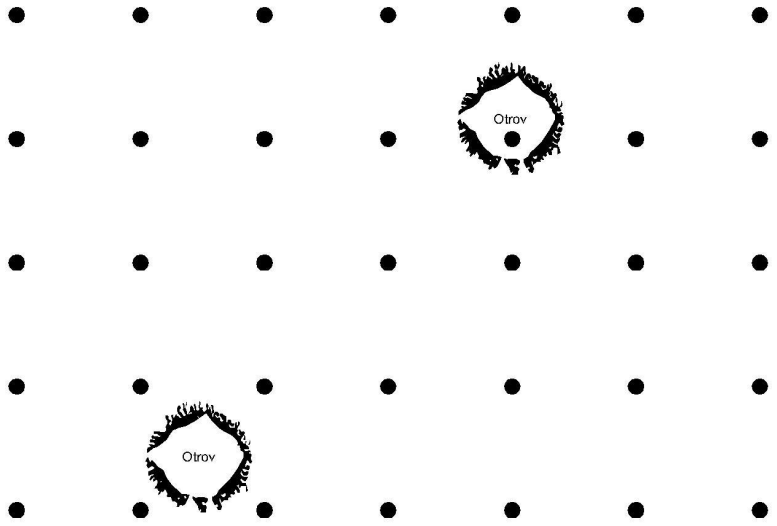
Primer 4.6 *Deponije u Sjedinjenim Američkim Državama*

Mnoge deponije u SAD sadrže otrovne materijale, ali se ne zna tačno gde su ovi materijali odloženi, da li su slučajno raspoređeni po deponiji, koncentrisani u jednoj oblasti ili ih uopšte nema. Praksa za otkrivanje ovih otrova je sledeća: bira se sistematski uzorak mreže tačaka sa kojih se uzima uzorak zemljišta i traži dokaz kontaminacije. Treba slučajno izabrati jednu tačku i konstruisati mrežu koja je sadrži tako da su tačke mreže na jednakoj međusobnoj udaljenosti.

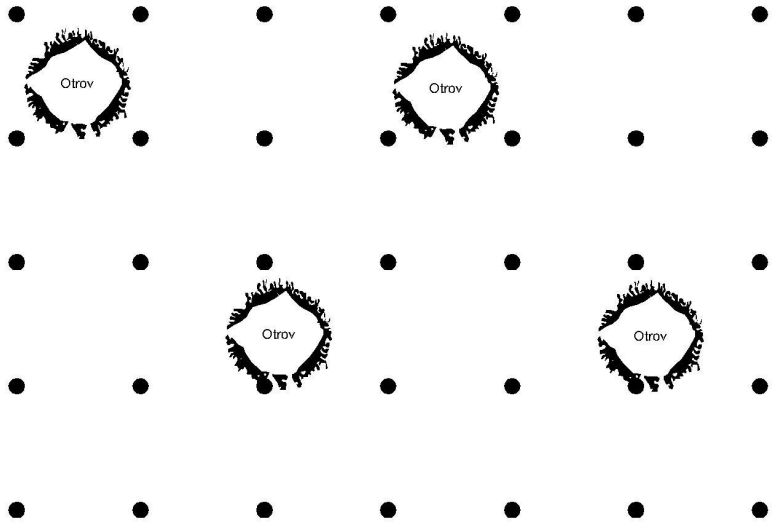
Prednosti uzimanja sistematskog uzorka u odnosu na prost slučajan uzorak su ravnomerna pokrivenost oblasti i lakše sprovođenje na terenu. U slučaju da nismo zabrinuti da su otrovni materijali distribuirani periodično ili imamo slabo predznanje o tome gde se oni mogu nalaziti, sistematski uzorak predstavlja prihvatljiv dizajn.

Postavljanjem mreže u sistematskom uzorkovanju može postojati bojazan da su otrovni materijali postavljeni tako da ih mreža neće detektovati. Ukoliko procenimo da postoji mogućnost da se to desi, preporučljivo je uzimanje stratifikovanog uzorka. Treba postaviti mrežu, birajući slučajnu tačku u svakom kvadratu sa koje će se uzimati uzorak zemljišta.

Mreža za detektovanje otrovnih materijala



Mreža za detektovanje otrovnih materijala - najgori mogući scenario. Pošto se otrovi javljaju sa sličnom periodičnošću kao i mreža, sistematski uzorak promašuje svaki otrov.



U slučaju periodičnosti, rešenje leži u korišćenju nekoliko sistematskih uzoraka, tj. umesto uzimanja jednog, može se uzimati više sistematskih uzoraka iz populacije. Tada se mogu koristiti formule za klaster uzorke za ocenu disperzije i svaki sistematski uzorak će se ponašati kao jedan klaster.

5

Primena klaster uzorka

U ovom odeljku predstavljena je primena klaster uzorka na realnom primeru. Za slučajan odabir klastera i uzorkovanih jedinica korišćen je Research Randomizer, dok su sva izračunavanja odrađena primenom Microsoft Excel-a.

Posmatran je spisak svih studenata Prirodno-matematičkog fakulteta u Novom Sadu (PMF) sa oznakom departmana, godine studija, osvojenim brojem poena u tekućoj školskoj godini i osvojenim ukupnim brojem poena tokom studija. Preko 95%-tnih intervala poverenja ocenjen je ukupan i prosečan broj osvojenih poena u tekućoj školskoj godini i ukupan i prosečan broj svih osvojenih poena u toku studija korišćenjem jednofaznog i dvofaznog klaster uzorka. U oba slučaja razmatrane su dve varijante, kada su klasteri:

- departmani PMF-a
- godine studija.

Prirodno-matematički fakultet u Novom Sadu u školskoj 2011/12. godini pohađalo je 3445 studenata raspoređenih na 4 godine studija i 5 departmana: Departmanu za biologiju i ekologiju (DBE), Departmanu za fiziku (DF), Departmanu za geografiju, turizam i hotelijerstvo (DGT), Departmanu za hemiju, biohemiju i zaštitu životne sredine (DH) i Departmanu za matematiku i informatiku (DMI).

U ovom primeru reč je o klasterima različitih veličina, jer departmani PMF-a imaju različit broj studenata, a i broj studenata PMF-a po godinama je, takođe, različit. S obzirom na to, populacione veličine ocenjene su na dva načina; korišćenjem nepristrasnog ocenjivača i ocenjivača količnika i dobijene ocene su upoređene.

Osim toga, ocene populacionih veličina dobijene korišćenjem jednofaznog i dvofaznog klaster uzorka poređene su sa ocenama dobijenim na osnovu prostog slučajnog uzorka istog obima.

U svim slučajevima, bilo da je reč o jednofaznom ili dvofaznom klaster uzorku, uzorkovana su $n = 2$ klastera.

Pošto je dostupan spisak svih studenata PMF-a i poena koje su oni osvojili u tekućoj školskoj godini, kao i ukupno osvojenih poena za vreme studiranja, u ovom primeru moguće je uraditi kompletan popis. Zahvaljujući tome, izračunate su stvarna ukupna i prosečna vrednost osvojenih poena u tekućoj školskoj godini i stvarna ukupna i prosečna vrednost osvojenih poena tokom studiranja. Dobijeni su podaci dati u sledećoj tabeli

	prosečan broj poena ($\bar{y}_U$)	ukupan broj poena (t)
Tekuća godina	28.33	98897.5
Na studijama	105.75	369189.5

5.1 Jednofazni klaster uzorak - klasteri su departmani

Od $N = 5$ departmana PMF-a slučajno su odabrana $n = 2$ departmana, korišćenjem Research Randomizer-a i to su DF i DH. Broj studenata na DF je $M_1 = 231$, a na DH je $M_2 = 570$. Pošto je reč o jednofaznom klaster uzorku, posmatrani su svi studenti unutar izabranih departmana, dakle $M_i = m_i$, $i = 1, 2$. Populacione veličine su ocenjene na dva načina, korišćenjem nepristrasne ocene i ocene količnika.

Nepristrasna ocena

Kao što je pomenuto kod jednofaznog klaster uzorka, nepristrasni ocenjivač za ukupnu vrednost populacije je

$$\hat{t}_{unb} = \frac{N}{n} \sum_{i \in \mathcal{S}} t_i,$$

sa standardnom greškom

$$SE(\hat{t}_{unb}) = N \sqrt{(1 - \frac{n}{N}) \frac{s_t^2}{n}},$$

gde je

$$s_t^2 = \frac{1}{n-1} \sum_{i \in \mathcal{S}} (t_i - \frac{\hat{t}_{unb}}{N})^2.$$

Nepristrasni ocenjivač srednje vrednosti populacije je

$$\hat{\bar{y}}_{unb} = \frac{\hat{t}_{unb}}{K},$$

a standardna greška je oblika

$$SE(\hat{\bar{y}}_{unb}) = \frac{SE(\hat{t}_{unb})}{K},$$

pri čemu je ukupan broj elemenata populacije, tj. studenata PMF-a

$$K = \sum_{i=1}^{5} M_i = 3445.$$

Korišćenjem prethodnih formula dobijene su sledeće ocene

	$\hat{\bar{y}}_{unb}$	$SE(\hat{\bar{y}}_{unb})$	$\hat{t}_{unb}$	$SE(\hat{t}_{unb})$
Tekuća godina	19.00	8.64	65447.5	29769.69
Na studijama	61.98	25.30	213530	87165.36

95%-tni intervali poverenja za ukupan i prosečan broj osvojenih poena konstruisani su na sledeći način

$$\hat{t}_{unb} \pm 1.96 SE(\hat{t}_{unb})$$

$$\hat{\bar{y}}_{unb} \pm 1.96 SE(\hat{\bar{y}}_{unb}).$$

Za konkretne podatke u našem primeru dobijeni su sledeći intervali poverenja:

Prosečan broj poena

	min	max
Tekuća godina	2.06	35.94
Na studijama	12.39	111.57

Ukupan broj poena

	min	max
Tekuća godina	7098.9	123796.1
Na studijama	42685.9	384374.1

Stvarne vrednosti prosečnog i ukupnog broja osvojenih poena dobijene popisom pripadaju konstruisanim intervalima poverenja, ali se može uočiti da su ocene prosečnog i ukupnog broja osvojenih poena, kako u tekućoj školskoj godini tako i tokom studiranja, dosta udaljene od stvarnih vrednosti ovih veličina dobijenih na osnovu ispitivanja čitave populacije. Takođe, dobijeni intervali poverenja su dosta široki usled veoma velikih standardnih grešaka.

Ocena količnika

Ukupna i srednja vrednost populacije se mogu oceniti i primenom ocenjivača količnika. U tom slučaju je ocenjivač srednje vrednosti populacije

$$\hat{\bar{y}}_r = \frac{\sum_{i\in\mathcal{S}} t_i}{\sum_{i\in\mathcal{S}} M_i},$$

ocenjivač ukupne vrednosti populacije je

$$\hat{t}_r = K\hat{\bar{y}}_r,$$

a formule za odgovarajuće standardne greške su oblika

$$SE(\hat{\bar{y}}_r) = \sqrt{(1-\frac{n}{N})\frac{1}{n\bar{M}_U^2}\frac{\sum_{i\in\mathcal{S}} M_i^2(\bar{y}_i-\hat{\bar{y}}_r)^2}{n-1}},$$

$$SE(\hat{t}_r) = N\sqrt{(1-\frac{n}{N})\frac{1}{n}\frac{\sum_{i\in\mathcal{S}} M_i^2(\bar{y}_i-\hat{\bar{y}}_r)^2}{n-1}},$$

gde je

$$\bar{M}_U = \frac{K}{N}.$$

Korišćenjem ovih ocenjivača dobijene su sledeće ocene za ukupnu i srednju vrednost populacije

	$\hat{\bar{y}}_r$	$SE(\hat{\bar{y}}_r)$	$\hat{t}_r$	$SE(\hat{t}_r)$
Tekuća godina	32.68	2.41	112592.58	8314.32
Na studijama	106.63	4.98	367346.24	17164.77

kao i odgovarajući intervali poverenja:

Prosečan broj poena

	min	max
Tekuća godina	27.95	37.41
Na studijama	96.87	116.40

Ukupan broj poena

	min	max
Tekuća godina	96296.5	128888.7
Na studijama	333703.3	400989.2

Za razliku od nepristrasne ocene, u ovom primeru, ocena količnika daje bolje rezultate. Stvarne vrednosti prosečnog i ukupnog broja poena u tekućoj godini kao i tokom studija pripadaju datim intervalima poverenja i oni su sada znatno uži usled manjih standardnih grešaka. Disperzija ocenjivača $\hat{t}_r$ zavisi od varijabilnosti sredina klastera i u ovom primeru je manja od disperzije nepristrasnog ocenjivača $\hat{t}_{unb}$ pa je ocena količnika preciznija od nepristrasne ocene. Dakle, na osnovu ovih rezultata može se potvrditi da je u slučaju kada se veličine klastera jako razlikuju, mnogo bolje koristiti ocenu količnika.

U ovom primeru je broj SSU u uzorku $M_1 + M_2 = 231 + 570 = 801$. Sada će ocene dobijene iz jednofaznog klaster uzorka biti upoređene sa ocenama dobijenim primenom prostog slučajnog uzorka istog obima.

Prost slučajan uzorak

Izabran je prost slučajan uzorak obima $n = 801$ od $N = 3445$ studenata. Određena je srednja vrednost uzorka $\bar{y}$ i uzoračka disperzija s^2 korišćenjem ugrađenih funkcija average i var u Excel-u. Standardna greška srednje vrednosti populacije je

$$SE(\bar{y}) = \sqrt{(1 - \frac{n}{N})\frac{s^2}{n}},$$

a nepristrasni ocenjivač ukupne vrednosti populacije je

$$\hat{t} = N\bar{y},$$

sa standardnom greškom

$$SE(\hat{t}) = \sqrt{N^2(1 - \frac{n}{N})\frac{s^2}{n}}.$$

95%-tni intervali poverenja za ukupnu i srednju vrednost populacije su oblika

$$\bar{y} \pm 1.96SE(\bar{y}),$$

$$\hat{t} \pm 1.96 SE(\hat{t}).$$

Korišćenjem prethodnih formula, dobijene su sledeće ocene

	$\bar{y}$	$SE(\bar{y})$	$\hat{t}$	$SE(\hat{t})$
Tekuća godina	28.57	0.56	98436.32	1944.47
Na studijama	108.70	2.06	374478.88	7101.79

i odgovarajući intervali poverenja:

Prosečan broj poena

	min	max
Tekuća godina	27.47	29.68
Na studijama	104.66	112.74

Ukupan broj poena

	min	max
Tekuća godina	94625.2	102247.5
Na studijama	360559.4	388398.4

Kao što je i očekivano, na osnovu dobijenih ocena i intervala poverenja, uočava se da je, u ovom primeru, prosto slučajno uzorkovanje efikasnije od jednofaznog klaster uzorkovanja.

5.2 Jednofazni klaster uzorak - klasteri su godine studija

Sada kao klastere posmatramo godine studija. Od $N = 4$ klastera slučajnim odabirom izabrana su $n = 2$ klastera i to su 1. i 2. godina. Prvu godinu pohađa $M_1 = 1089$, a drugu $M_2 = 860$ studenata.

Nepristrasna ocena

Korišćenjem nepristrasnih ocenjivača, dobijene su sledeće ocene za ukupnu i srednju vrednost populacije

	$\hat{\bar{y}}_{unb}$	$SE(\hat{\bar{y}}_{unb})$	$\hat{t}_{unb}$	$SE(\hat{t}_{unb})$
Tekuća godina	32.11	3.89	110614	13399.67
Na studijama	65.54	14.66	225782	50517.12

a odgovarajući 95%-tni intervali poverenja su oblika:

Prosečan broj poena

	min	max
Tekuća godina	24.48	39.73
Na studijama	36.80	94.28

Ukupan broj poena

	min	max
Tekuća godina	84350.6	136877.4
Na studijama	126768.4	324795.6

Može se uočiti da su ocene prosečnog broja osvojenih poena u toku studija $\hat{\bar{y}}_{unb} = 65.54$ i ukupnog broja osvojenih poena u toku studija $\hat{t}_{unb} = 225782$ potcenjene u odnosu na stvarne vrednosti, $\bar{y}_U = 105.75$ i $t = 369189.5$, dobijene ispitivanjem čitave populacije, koje u ovom konkretnom primeru ne pripadaju dobijenim intervalima poverenja. To je i bilo očekivano pošto su, slučajnim odabirom, uzorak činile prva i druga godina na kojima studenti imaju najmanji ukupan broj poena u toku studija.

Ocena količnika

Sada su navedene ocene ukupne i srednje vrednosti populacije dobijene korišćenjem ocenjivača količnika

	$\hat{\bar{y}}_r$	$SE(\hat{\bar{y}}_r)$	$\hat{t}_r$	$SE(\hat{t}_r)$
Tekuća godina	28.38	1.22	97759.17	4209.61
Na studijama	57.92	20.11	199543.10	69275.62

i odgovarajući intervali poverenja su:

Prosečan broj poena

	min	max
Tekuća godina	25.98	30.77
Na studijama	18.51	97.34

Ukupan broj poena

	min	max
Tekuća godina	89508.3	106010.0
Na studijama	63762.9	335323.3

Ponovo se mogu uočiti potcenjene ocene prosečnog i ukupnog broja osvojenih poena u toku studija i intervali poverenja koji ne sadrže stvarne vrednosti ovih veličina, pa se dolazi do zaključka da, poput nepristrasne ocene, ni ocena količnika u ovom slučaju ne daje dovoljno dobre rezultate za osvojene poene u toku studija. Međutim, ocena ukupnog i prosečnog broja osvojenih poena u tekućoj godini dobijena na ovaj način preko količnika je preciznija od nepristrasne ocene.

5.3 Dvofazni klaster uzorak - klasteri su departmani

U ovom delu predstavljeni su rezultati dobijeni primenom dvofaznog klaster uzorka. Sada ponovo, departmani predstavljaju klastere. Radi mogućnosti poređenja sa jednofaznim klaster uzorkom koji su činili DF i DH, dvofazni klaster uzorak je uzet na istim departmanima. Već je rečeno da DF ima $M_1 = 231$ studenata, dok DH ima $M_2 = 570$ studenata.

Ukupan broj SSU u uzorku je određen tako da čini otprilike 10% od ukupnog broja elemenata populacije, tj. $m_1 + m_2 = 350$, dok se m_1 i m_2 određuju na sledeći način

$$m_i = M_i \frac{m_1 + m_2}{M_1 + M_2} = M_i \frac{350}{801}, \quad i = 1, 2.$$

Vrednosti m_1 i m_2 se zaokružuju na ceo broj, pa se u ovom slučaju dobija $m_1 = 101$ i $m_2 = 249$.

Nepristrasna ocena

Kod dvofaznog klaster uzorka nepristrasni ocenjivač ukupne vrednosti populacije i odgovarajuća standardna greška su oblika

$$\hat{t}_{unb} = \frac{N}{n} \sum_{i \in \mathcal{S}} M_i \bar{y}_i,$$

$$SE(\hat{t}_{unb}) = \sqrt{N^2 (1 - \frac{n}{N}) \frac{s_t^2}{n} + \frac{N}{n} \sum_{i \in \mathcal{S}} (1 - \frac{m_i}{M_i}) M_i^2 \frac{s_i^2}{m_i}},$$

pri čemu je

$$s_t^2 = \frac{\sum_{i \in \mathcal{S}} (\hat{t}_i - \frac{\hat{t}_{unb}}{N})^2}{n - 1},$$

$$s_i^2 = \frac{\sum_{j \in \mathcal{S}_i} (y_{ij} - \bar{y}_i)^2}{m_i - 1},$$

dok su ocenjivač srednje vrednosti populacije i njegova standardna greška oblika

$$\hat{\bar{y}}_{unb} = \frac{\hat{t}_{unb}}{K},$$

$$SE(\hat{\bar{y}}_{unb}) = \frac{SE(\hat{t}_{unb})}{K}.$$

Korišćenjem prethodnih formula dobijene su sledeće ocene

	$\hat{\bar{y}}_{unb}$	$SE(\hat{\bar{y}}_{unb})$	$\hat{t}_{unb}$	$SE(\hat{t}_{unb})$
Tekuća godina	18.33	8.55	63163.85	29454.12
Na studijama	60.44	23.56	208198.70	81180.15

i intervali poverenja:

Prosečan broj poena

	min	max
Tekuća godina	1.58	35.09
Na studijama	14.25	106.62

Ukupan broj poena

	min	max
Tekuća godina	5433.8	120893.9
Na studijama	49085.6	367311.8

Ocena količnika

Kod dvofaznog klaster uzorka, ukupna i srednja vrednost populacije se mogu oceniti i primenom ocenjivača količnika. Tada je ocenjivač srednje vrednosti populacije

$$\hat{\bar{y}}_r = \frac{\sum_{i \in \mathcal{S}} M_i \bar{y}_i}{\sum_{i \in \mathcal{S}} M_i},$$

dok je standardna greška

$$SE(\hat{\bar{y}}_r) = \sqrt{\frac{1}{\bar{M}^2}[(1 - \frac{n}{N})\frac{s_r^2}{n} + \frac{1}{nN}\sum_{i \in \mathcal{S}} M_i^2 (1 - \frac{m_i}{M_i})\frac{s_i^2}{m_i}]},$$

gde je

$$s_r^2 = \frac{\sum_{i\in\mathcal{S}}(M_i\bar{y}_i - M_i\hat{\bar{y}}_r)^2}{n-1},$$

$$s_i^2 = \frac{\sum_{j\in\mathcal{S}_i}(y_{ij} - \bar{y}_i)^2}{m_i - 1},$$

a $\bar{M}$ prosečna veličina klastera. Ocenjivač ukupne vrednosti populacije je

$$\hat{t}_r = K\hat{\bar{y}}_r,$$

a standardna greška je

$$SE(\hat{t}_r) = K\,SE(\hat{\bar{y}}_r).$$

Na osnovu formula, dobijene su sledeće ocene

	$\hat{\bar{y}}_r$	$SE(\hat{\bar{y}}_r)$	$\hat{t}_r$	$SE(\hat{t}_r)$
Tekuća godina	31.54	2.55	108663.90	8783.39
Na studijama	103.97	3.85	358174.55	13272.61

a 95%-tni intervali poverenja za ukupan i prosečan broj osvojenih poena u toku školske godine, kao i u toku studija izgledaju ovako:

Prosečan broj poena

	min	max
Tekuća godina	26.55	36.54
Na studijama	96.42	111.52

Ukupan broj poena

	min	max
Tekuća godina	91448.5	125879.3
Na studijama	332160.2	384188.9

Kao i kod jednofaznog klaster uzorka, nameće se zaključak da je ocena količnika preciznija od nepristrasne ocene kada su klasteri različitih veličina, kao što je slučaj u ovom primeru.

Upoređivanjem rezultata dobijenih na osnovu dvofaznog klaster uzorka sa rezultatima dobijenim na osnovu jednofaznog klaster uzorka u slučaju kada uzorak čine klasteri DF i DH, uočava se velika sličnost, što znači da uzorkovanje svih elemenata klastera kod jednofaznog klaster uzorka nije dalo više informacija od uzorkovanja samo pojedinih elemenata kod dvofaznog klaster uzorka.

Sada je ponovo uzet dvofazni klaster uzorak u slučaju kada su klasteri departmani PMF-a, ali su slučajno izabrani klasteri DF i DGT. DF ima $M_1 = 231$ studenata, a DGT $M_2 = 1072$ studenata. Zahtevan broj sekundarnih uzoračkih jedinica je ponovo 350, dok je sada $m_1 = 62$ i $m_2 = 288$.

Nepristrasna ocena

Koristeći već navedene formule za nepristrasne ocenjivače kod dvofaznog klaster uzorka, dobijene su ocene

	$\hat{\bar{y}}_{unb}$	$SE(\hat{\bar{y}}_{unb})$	$\hat{t}_{unb}$	$SE(\hat{t}_{unb})$
Tekuća godina	26.53	14.84	91411.42	51111.07
Na studijama	104.39	58.95	359632.20	203095.67

i odgovarajući intervali poverenja:

Prosečan broj poena

	min	max
Tekuća godina	-2.54	55.61
Na studijama	-11.16	219.94

Ukupan broj poena

	min	max
Tekuća godina	-8766.3	191589.1
Na studijama	-38435.3	757699.7

Uočavaju se negativne donje granice intervala poverenja. To se dešava zbog izuzetno velikih standardnih grešaka što je u ovom primeru očekivano jer se veličine izabranih klastera veoma razlikuju, a u tom slučaju nepristrasne ocene nisu dobre.

Ocena količnika

Korišćenjem ocenjivača količnika kod dvofaznog klaster uzorka dobijene su ocene

	$\hat{\bar{y}}_r$	$SE(\hat{\bar{y}}_r)$	$\hat{t}_r$	$SE(\hat{t}_r)$
Tekuća godina	28.06	1.64	96673.01	5656.71
Na studijama	110.40	6.94	380332.44	23915,61

i 95%-tni intervali poverenja:

Prosečan broj poena

	min	max
Tekuća godina	24.84	31.28
Na studijama	96.79	124.01

Ukupan broj poena

	min	max
Tekuća godina	85585.9	107760.2
Na studijama	333457.9	427207.0

Za razliku od nepristrasne ocene, ocena količnika je pravi izbor u ovom slučaju. Intervali poverenja sadrže stvarne vrednosti populacionih veličina, a standardne greške su neuporedivo manje nego što je bio slučaj kod nepristrasnih ocena. Time su potvrđeni teorijski rezultati da u slučaju kada se veličine klastera jako razlikuju, treba koristiti ocenu količnika jer je ona mnogo preciznija od nepristrasne ocene.

5.4 Dvofazni klaster uzorak - klasteri su godine studija

Slučajnim odabirom, određeno je da uzorak čine 3. i 4. godina studija koju, respektivno, pohađa $M_1 = 722$ i $M_2 = 774$ studenta. Broj sekundarnih uzoračkih jedinica je 350. Sa treće godine studija uzorkovana su $m_1 = 169$ studenata, a sa četvrte $m_2 = 181$ studenata.

Nepristrasna ocena

Primenjujući iste formule kao u slučaju kada su klasteri bili departmani, dobijene su sledeće ocene

	$\hat{\bar{y}}_{unb}$	$SE(\hat{\bar{y}}_{unb})$	$\hat{t}_{unb}$	$SE(\hat{t}_{unb})$
Tekuća godina	24.23	1.47	83472.27	5052.14
Na studijama	146.26	19.64	503855.61	67644.22

a odgovarajući intervali poverenja su:

Prosečan broj poena

	min	max
Tekuća godina	21.36	27.10
Na studijama	107.77	184.74

Ukupan broj poena

	min	max
Tekuća godina	73570.1	93374.5
Na studijama	371272.9	636438.3

Nepristrasne ocene za prosečan i ukupan broj poena osvojenih tokom studija su precenjene, a stvarne vrednosti prosečnog i ukupnog broja poena osvojenih tokom studija dobijene popisom ne pripadaju dobijenim intervalima poverenja. Razlog za to je odabir treće i četvrte godine studija u uzorak jer su upravo te dve godine studija one na kojima studenti imaju najveći broj poena.

Ocena količnika

Korišćenjem ocenjivača količnika dobijene su sledeće ocene

	$\hat{\bar{y}}_r$	$SE(\hat{\bar{y}}_r)$	$\hat{t}_r$	$SE(\hat{t}_r)$
Tekuća godina	27.90	2.03	96110.28	7002.79
Na studijama	168.40	16.04	580141.24	55277.51

pri čemu se, takođe, uočava precenjenost ocena prosečnog i ukupnog broja poena tokom studija u odnosu na stvarne vrednosti.
Intervali poverenja su:

Prosečan broj poena

	min	max
Tekuća godina	23.91	31.88
Na studijama	136.95	199.85

Ukupan broj poena

	min	max
Tekuća godina	82384.8	109835.8
Na studijama	471797.3	688485.2

Pošto su izabrani klasteri sličnih veličina, ocene količnika i nepristrasne ocene su slične preciznosti.

Sada će ocene dobijene primenom dvofaznog klaster uzorka biti upoređene sa ocenama dobijenim na osnovu prostog slučajnog uzorka istog obima. Dakle, uzet je prost slučajan uzorak obima $n = 350$ od $N = 3445$ studenata.

Korišćenjem formula za prost slučajan uzorak, dobijene su nepristrasne ocene

	$\bar{y}$	$SE(\bar{y})$	$\hat{t}$	$SE(\hat{t})$
Tekuća godina	27.64	0.95	95222.57	3254.95
Na studijama	97.78	3.14	336857.64	10819.10

a odgovarajući 95%-tni intervali poverenja su:

Prosečan broj poena

	min	max
Tekuća godina	25.79	29.49
Na studijama	91.63	103.94

Ukupan broj poena

	min	max
Tekuća godina	88842.9	101602.3
Na studijama	315652.2	358063.1

Kao što je i očekivano, uočava se da su ocene dobijene na osnovu prostog slučajnog uzorka preciznije od ocena dobijenih primenom dvofaznog klaster uzorka istog obima, upravo iz razloga što su studenti istog departmana (godine studija) sličniji nego slučajno odabrani studenti sa celog fakulteta.

6

Zaključak

U ovom radu razmatrano je klaster uzorkovanje sa jednakim verovatnoćama. Posvećena je pažnja ocenjivanju populacionih veličina kako u jednofaznom tako i u dvofaznom klaster uzorku sa akcentom na nepristrasnim ocenama i ocenama količnika. Osim toga, prikazan je i sistematski uzorak koji predstavlja specijalan slučaj klaster uzorka.

Jednofazni klaster uzorak se koristi kada su troškovi uzorkovanja sekundarnih uzoračkih jedinica zanemarljivi u poređenju sa troškovima uzorkovanja klastera. U jednofaznom klaster uzorku ispituju se svi elementi unutar izabranog klastera. Ovi elementi mogu imati veoma slične osobine, pa je ispitivanje svih njih ponekad skupo i nepotrebno. U takvim situacijama jeftinije je biranje poduzorka iz svakog izabranog klastera i tada je reč o dvofaznom klaster uzorku.

Osnovni razlog za primenu klaster uzorka prilikom ispitivanja populacije je ekonomičnost. Osim toga, klaster uzorkovanje je ponekad jedina izvodljiva metoda verovatnosnog uzorkovanja pošto uzoračke okvire ciljnih populacija često čine liste klastera. To je posebno prisutno u istraživanju ljudske populacije kada su jedini dostupni podaci o populaciji spiskovi domaćinstava.

Na osnovu teorijske analize i analize na konkretnom realnom primeru, može se zaključiti da su ocene populacionih veličina dobijene klaster uzorkovanjem u opštem slučaju manje precizne od ocena dobijenih prostim slučajnim uzorkovanjem, pošto su često elementi istog klastera sličniji nego slučajno odabrani elementi iz cele populacije. Uzorkovanjem sličnih elemenata iz jednog klastera ne dobijaju se nove informacije, pa to smanjuje preciznost ocena populacionih parametara. I pored toga, klaster uzorci se opravdano veoma često koriste i imaju veliku primenu u praksi kada su u pitanju velika istraživanja, jer su ekonomičniji i jednostavniji od nekih drugih tipova uzoraka, a mogu pružiti precizne i korisne rezultate.

Literatura

[1] William Gemmell Cochrain, *Sampling Techniques*, Wiley (1977)

[2] Ljiljana Cvetković, *Poslovna statistika*, Futura publikacije, Novi Sad (2006)

[3] Johnnie Daniel, *Sampling Essentials: Practical Guidelines for Making Sampling Choices*, SAGE Publications (2012)

[4] Gary T. Henry, *Practical sampling*, SAGE Publications (1990)

[5] Leslie Kish, *Survey Sampling*, Wiley (1995)

[6] Sharon L. Lohr, *Sampling: Design and Analysis*, Duxbury Press (1999), 2^{nd} edition (2010)

[7] Mankal N. Murthy *Sampling: theory and methods*, Statistical Pub. Society (1967)

[8] Ljiljana Petrović, *Teorija uzoraka i planiranje eksperimenata*, Ekonomski fakultet, Univerzitet u Beogradu (2008)

[9] C. E. Sarndal, B. Swenson & J. Wretman *Model assisted survey sampling* New York, NY: Springer - Verlag (1992)

[10] Steven K. Thompson, *Sampling*, Wiley, 1^{st} edition (1992), 2^{nd} edition (2012)

[11] Steven K. Thompson, *Stratified adaptive cluster sampling*, Wiley, 3^{rd} edition (2012)

[12] http://personal.georgiasouthern.edu/∼rvogel/PDFs/Chapter%204.pdf

[13] http://personal.georgiasouthern.edu/∼rvogel/PDFs/Chapter%207.pdf

[14] http://personal.georgiasouthern.edu/∼rvogel/PDFs/Chapter%208.pdf

[15] http://www.stat.purdue.edu/∼jennings/stat522/notes/topic5.pdf

KRATKA BIOGRAFIJA

Dunja Arsić je rođena 28. oktobra 1988. godine u Novom Sadu. Osnovnu školu "Kosta Trifković" završila je 2003. godine sa odličnim uspehom i Vukovom diplomom. Iste godine upisuje prirodno-matematički smer Gimnazije "Isidora Sekulić" u Novom Sadu, koju takođe završava kao nosilac Vukove diplome. Posle završetka gimnazije, 2007. godine, upisuje osnovne akademske studije na Prirodno-matematičkom fakultetu, Departman za matematiku i informatiku, smer matematika finansija. Diplomirala je 28. juna 2011. godine sa prosečnom ocenom 10.00.

Odmah zatim, nastavlja školovanje na istom fakultetu, upisujući master studije primenjene matematike, smer matematika finansija. Položila je sve ispite predviđene nastavnim planom i programom master studija, zaključno sa junskim ispitnim rokom i tako ostvarila pravo na odbranu master rada. Stipendista je Fonda za mlade talente Republike Srbije kako na osnovnim tako i na master studijama.

U Novom Sadu, 2012.

UNIVERZITET U NOVOM SADU

PRIRODNO - MATEMATIČKI FAKULTET

KLJUČNA DOKUMENTACIJSKA INFORMACIJA

Redni broj:
RBR

Identifikacioni broj:
IBR

Tip dokumentacije: Monografska dokumentacija
TD

Tip zapisa: Tekstualni štampani materijal
TZ

Vrsta rada: Master rad
VR

Autor: Dunja Arsić
AU

Mentor: Prof. dr Sanja Rapajić
MN

Naslov rada: Klaster uzorci sa jednakim verovatnoćama
NR

Jezik publikacije: srpski (latinica)
JP

Jezik izvoda: s / e
JI

Zemlja publikovanja: Srbija
ZP

Uže geografsko područje: Vojvodina
UGP

Godina: 2012
GO

Izdavač: Autorski reprint
IZ

Mesto i adresa: Novi Sad, Departman za matematiku i informatiku, Prirodno-matematički fakultet, Trg Dositeja Obradovića 4
MA

Fizički opis rada: (6, 78, 15, 43, 4, 3, 0)
FO

Naučna oblast: Matematika
NO

Naučna disciplina: Primenjena statistika
ND

Ključne reči: klaster uzorkovanje, jednofazni klaster uzorak, dvofazni klaster uzorak, nepristrasne ocene, ocene količnika, prost slučajan uzorak
PO

UDK:

Čuva se:
ČU

Važna napomena:
VN

Izvod:
IZ

U radu je razmatran klaster uzorak, koji uz prost slučajan uzorak i stratifikovan uzorak spada u osnovne tipove verovatnosnih uzoraka. Predstavljeni su i analizirani ocenjivači ukupne i srednje vrednosti populacije za jednofazni i dvofazni klaster uzorak, sa akcentom na nepristrasnim ocenama i ocenama količnika. Osim toga, prikazan je i sistematski uzorak kao specijalan slučaj klaster uzorka.

Datum prihvatanja teme od strane NN veća: 10.05.2012.
DP

Datum odbrane:
DO

Članovi komisije:
KO

Predsednik:	Dr Zorana Lužanin, redovni profesor Prirodno-matematičkog fakulteta u Novom Sadu
Mentor:	Dr Sanja Rapajić, vanredni profesor Prirodno-matematičkog fakulteta u Novom Sadu
Član:	Dr Sanja Konjik, docent Prirodno-matematičkog fakulteta u Novom Sadu

UNIVERSITY OF NOVI SAD

FACULTY OF SCIENCES

KEY WORDS DOCUMENTATION

Accession number:
ANO

Identification number:
INO

Document type: Monograph type
DT

Type of record: Printed text
TR

Contents Code: Master's thesis
CC

Author: Dunja Arsić
AU

Mentor: Prof. dr Sanja Rapajić
MN

Title: Cluster samples with equal probabilities
TI

Language of text: Serbian
LT

Language of abstract: English
LA

Country of publication: Serbia
CP

Locality of publication: Vojvodina
LP

Publication year: 2012
PY

Publisher: Author's reprint
PU

Publ. place: Novi Sad, Department of Mathematics and Informatics, Faculty of Sciences, Trg Dositeja Obradovića 4
PP

Physical description: (6, 78, 15, 43, 4, 3, 0)
PD

Scientific field: Mathematics
SF

Scientific discipline: Applied statistics
SD

Key words: cluster sampling, one-stage cluster sample, two-stage cluster sample, unbiased estimation, ratio estimation, simple random sample
SKW

UC:

Holding data:
HD

Note:
N

Abstract:
AB

Cluster sample is proposed in this master thesis. It is one of the main types of probability samples along with simple random sample and stratified sample. The estimators of population total and mean for both one-stage and two-stage cluster sample are presented and analyzed, with an emphasis on unbiased estimation and ratio estimation. Systematic sample, as a special case of cluster sample, is also proposed.

Accepted by the Scientific Board on: 10.05.2012.
ASB

Defended:
DE

Thesis defend board:
DB

President:	Dr Zorana Lužanin, Full Professor, Faculty of Sciences, University of Novi Sad
Mentor:	Dr Sanja Rapajić, Associate Professor, Faculty of Sciences, University of Novi Sad
Member:	Dr Sanja Konjik, Assistant Professor, Faculty of Sciences, University of Novi Sad