Skripta iz Uvoda u teoriju uzoraka

12. mart 2020

Sadržaj

1 nedelja

1.1 Naučno istraživanje

Naučno istraživanje je sistematsko, plansko i objektivno ispitivanje nekog problema, prema određenim metodološkim pravilima, čija je svrha da se pruži pouzdan i precizan odgovor na unapred postavljeno pitanje.

Može se shvatiti kao kritički, kontrolisani i ponovljivi proces sticanja novih znanja, neophodnih (a ponekad i dovoljnih) za identifikovanje, određivanje i rešavanje naučnih (teorijskih i empirijskih) problema.

Teorijsko istraživanje vs Empirijsko (**iskustveno**) istraživanje.

Svako naučno istraživanje ima više međusobno logično povezanih faza.
Faze su:

- identifikovanje i određivanje problema
- određivanje cilja istraživanja
- definisanje ključnih izraza
- postavljanje hipoteze i izvođenje logičkih posledica iz hipoteze
- izbor istraživačke strategije i plana istraživanja
- razvijanje mernih i drugih instrumenata istraživanja
- određivanje onovnog skupa (populacije) i odabir uzorka
- sprovođenje istraživanja i prikupljanje relevantnih podataka
- obrađivanje i analiza podataka dobijenih istraživanjem
- tumačenje rezultata istraživanja i izvođenje zaključ(a)ka
- izrada izveštaja o obavljenom istraživanju
- prezentacija rezultata istraživanja

1.2 Osnovni pojmovi

Entitet/**jedinica posmatranja** (en. 'observation unit') - živo biće ili objekat čija su svojstva predmet istraživanja.

Populacija ('population') - skup / kolekcija entiteta.

Na osnovu broja entiteta, tj. **obima** / veličine **populacije** ('populationsize') N, može biti:

- konačna populacija$-N$ je prirodan broj
- beskonačna populacija$-N\rightarrow+\infty$

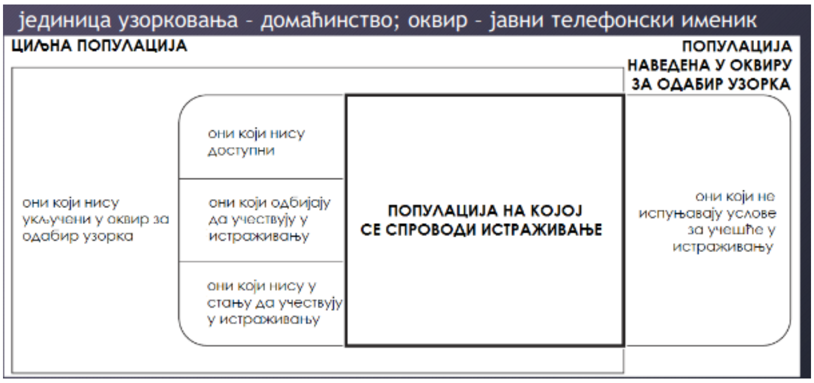
Trebalo bi razlikovati:

- ciljnu populaciju('target population')
- populaciju na kojoj se efektivno sprovodi istraživanje ('study population') [1]

Uzorak ('sample') - podskup populacije; sadrži izvesne entitete koji potiču iz populacije, na bazi čijeg proučavanja će se izvoditi zaključci o čitavoj populaciji
Obim uzorka ('samplesize') n [2]
Jedinica uzorkovanja ('sampling unit') [3]
Okvir za odabir uzorka ('sampling frame') - popis (ili neka druga specifikacija) svih jedinica uzorkovanja

Npr. svakoj jedinici uzorkovanja pridruži se različit prirodan broj (počevši od 1). Ti brojevi nazivaju se **oznake jedinica**, služe za njihovo identifikovanje i ostaju nepromenjeni sve do kraja istraživanja.

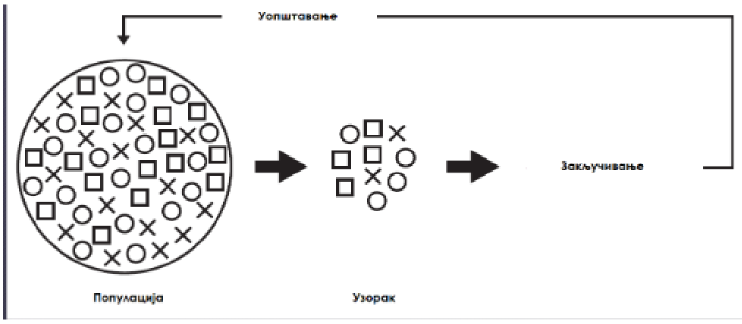
Primer - telefonsko istraživanje biračkog tela.



Zašto uzorkovanje?
Potpuno ispitivanje populacije (proučavanje tzv. **cenzusa**) je, u mnogim slučajevima, neracionalno ili čak principijelno nemoguće. Čak i onda kada postoji mogućnost potpunog ispitivanja populacije istraživač se obično opredeljuje za **delimično** ispitivanje (proučavanje uzorka) jer je (u odnosu na potpuno ispitivanje):

- jeftinije
- brže
- kontrola tačnosti prikuplenih podataka je jednostavnija i lakša

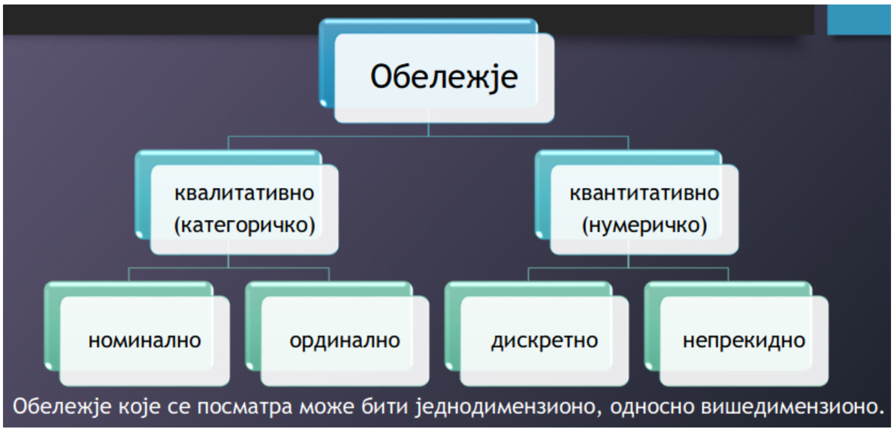
[1] Nadalje se pretpostavlja: target population = study population i $N < +\infty$
[2] Uvek konačna vrednost
[3] U opštem slučaju nije isto što i jedinica posmatranja, koja predstavlja osnovni objekat posmatranja i prikupljanja informacija. Jedinice uzorkovanja su međusobno disjunktni skupovi entiteta



Termin populacija odnosi se na skup entiteta istovrsnih u odnosu na jedno ili više zajedničkih svojstava, koja se mogu posmatrati. Ipak, entiteti, iako **istovrsni**, **nisu istovetni**.
Određivanje populacije predstavlja značajnu i, neretko, tešku fazu istraživanja. Populacija mora biti definisana: pojmovno (u smislu svog sadržaja - šta su entiteti, a šta jedinice uzorkovanja?), prostorno i vremenski.

Obeležje ('study variable') - posmatrana zajednička karakteristika svih entiteta u populaciji, tj. preciznije, izvesno varijabilno svojstvo od interesa, koje je određeno za svaki entitet u populaciji. [4]

1.3 Tipovi obeležja



[4] Obeležje najčešće nije neko od definicionih svojstava populacije.

Primer: tipovi obeležja

- kvalitativna
  - nominalna
    - boja očiju, krvna grupa
    - etnička / verska pripadnost
    - radna mesta na fakultetu
    - raspoloženje građana Srbije prema pristupanju u EU
    - posedovanje profila na društvenim mrežama
  - ordinalna
    - nivo akademskih studija
    - čin oficira u vojsci
    - ocena restorana na Tripadvisor
    - stanje pacijenta
    - intezitet bola
- kvantitativna
  - diskretna
    - broj stanovnik sa pravom glasa u određenoj opštini
    - broj blizanaca rođenih u toku godine u određenoj regiji
    - broj kućnih ljubimaca u domaćinstvu
  - neprekidna
    - visina, težina, starost, IQ
    - dužina lista određene biljne vrste
    - koncetracija soli u morskoj vodi

Primer: populacija i obeležje

- Populacija: skup studenata koji su upisali Uvod u teoriju uzoraka školske 2019/20. godine.
  Obeležje: pol; broj položenih ispita, broj položenih ESPB bodova,prosečna ocena svih položenih ispita –zaključno sa rokom Januar 2 ove školske godine; ocena na kursu Statistika

- Populacija: skup svih poljoprivrednih gazdinstava u Srbiji(referentni period–oktobar/novembar2018)
  Obeležje:površina korišćenog poljoprivrednog zemljišta; broj grla stoke; primenjeni proizvodni metodi

- Populacija: skup svih domaćinstava u regionu Šumadije i Istočne Srbije(referentni period –2017. g)
  Obeležje: lična potrošnja domaćinstva (mesečni prosek)

- Populacija: jedna serija LED sijalica izvesnog proivođača.
  Obeležje: dužina radnog veka sijalice u satima.

- Populacija: skup svih meseci u periodu od 2000. do 2016. g.
  Obeležje: mesečni broj vetrovitih dana u Vršcu

Obeležje se može shvatiti kao funkcija koja entitetima u populaciji pridružuje realne brojeve ili neke druge vrednosti.
Neka je data populacija sa N jedinica, koje su u okviru za odabiruzorka označene brojevima iz skupa $\omega = 1, 2, ..., N$ (i time jednoznačno određene)i neka je Y obeležje od interesa.Neka je sa y_k označena vrednost obeležja Y entiteta označenog sa k.

Zadatak pri istraživanju obično se svodi na donošenje zaključaka o (nepoznatoj) vrednosti realne funkcije

$$\theta = f(y_1, y_2, , y_n)$$

, koja se naziva **populacijska vrednost** ('population value') ili **parametar populacije**.

Najčešće funkcije koje se pojavljuju kao parametri populacije:

- Kvantitativna obeležja
  - populacijska srednja vrednost ('population mean')

    $$m_Y = m = \frac{1}{N}\sum_{k=1}^{N} y_K$$

  - populacijski total ('population total')

    $$\tau_Y = \tau = \sum_{k=1}^{N} y_K = N m_Y$$

  - populacijska disperzija ('population variance') / standardno odstupanje

    $$\sigma_Y^2 = \sigma^2 = \frac{1}{N-1}\sum_{k=1}^{N}(y_K - m_Y)^2$$

    i

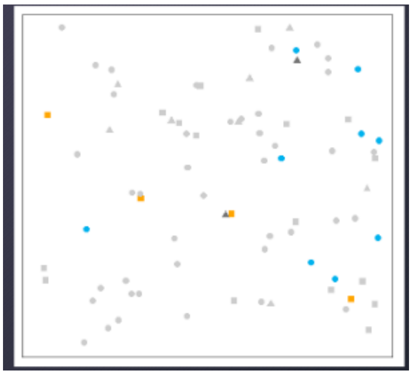
    $$\sigma_Y = \sigma = \sqrt{\sigma_Y^2}$$

- Kvalitativna obeležja
  - populacijska proporcija ('population proportion')
  - populacijska medijana, kvantili, moda...

Ideja je da se zaključci o populacijskim vrednostima donose na osnovu informacija dobijenih ispitivanjem uzorka.
„Dobar“ uzorak ima osobinu **reprezentativnosti**. To je uzorak koji predstavlja „umanjenu“, a nikako „iskrivljenu“, niti „uveličanu“ sliku jednog dela populacije. Uzorak sa ovom osobinom verno odslikava strukturu populacije koju predstavlja, „izgledajući“ kao i populacija u svim aspektima relevatnim za istraživanje.

Na reprezentativnost uzorka utiču:

- tip uzorka (prema metodu odabira)
- veličina uzorka
- varijabilnost posmatranog obeležja



Plan uzorkovanja ('sampling design') poseduje dve osnovne komponente:

- metod odabira uzorka
- metod zaključivanja

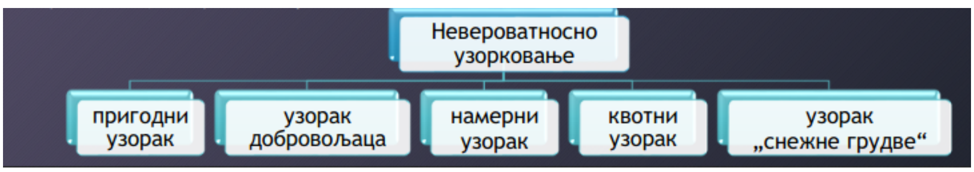
Metod odabira uzorka je postupak kojim se biraju elementi populacije u uzorak, uz određivanje adekvatnog obima uzorka.
Ovi metodi se mogu podeliti u dve grupe:

- **Verovatnosno uzorkovanje** ('probability sampling')
- **Neverovatnosno uzorkovanje** ('nonprobability sampling')

1.4 Nevereovatnosno uzorkovanje

Ovakvi metodi uzorkovanja ne zasnivaju se na teoriji verovatnoća, nego na određenim kriterijumima istraživača.
Dakle, njihova osnovna osobina jeste da se uzorkovanje vrši na osnovu **subjektivne procene istraživača**, a ne slučajnim izborom. Njima se pribegava onda kada je, zbog ograničenih vremenskih rokova, iznosa troškova i osetljivosti predmeta istraživanja (etičkih obzira), teško sprovesti slučajno uzorkovanje.

- Prednosti: efikasnije se primenjuju kod eksplorativnih istraživanja (pilot istraživanja, studije u cilju dokazivanja koncepta, kvalitativna istraživanja, studije za generisanje hipoteza), čiji cilj nije precizno zaključivanje o parametrima populacije na osnovu reprezentativnog uzorka.
- Mane: nije moguće određivanje kvaliteta uzorka, a samim tim ni kvantifikovanje tačnosti zaključivanja (zaključivanje je ovde analitičko).



1.5 Verovatnosno uzorkovanje

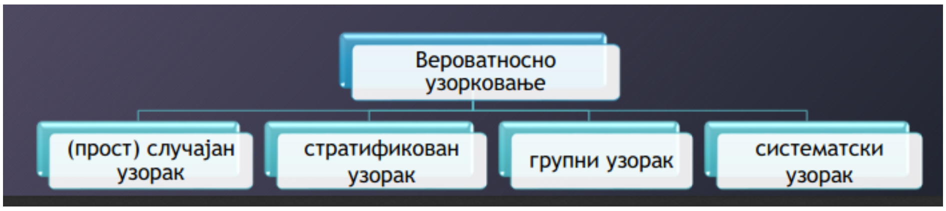
Ovakvi metodi uzorkovanja zasnivaju se na teoriji verovatnoća, tj. na „planiranoj“ slučajnosti. Mogući uzorci su faktički matematički konstruisani, i za svakog od njih poznata je verovatnoća da bude odabran. Dakle, uzorkovanje se vrši u skladu sa raspodelom verovatnoća, definisanom na kolekciji svih mogućih uzoraka.

•Neka je sa Ω označen skup oznaka jedinica u populaciji i neka $s \subset \Omega$ predstavlja uzorak. Verovatnosno uzorkovanje se zasniva na poznavanju raspodele verovatnoća $p(\bullet)$:

$$p(s) \geq 0, \forall s \subset \Omega, \sum_{s \subset \Omega} p(s) = 1$$

Slučajan uzorak S je onda slučajan skup oznaka jedinica sa raspodelom verovatnoća:

$$P\{S = s\} = p(s), \forall s \subset \Omega$$



Prednosti:

- doslednom primenom isključuje se postojanje bilo kakve pristrasnosti, što doprinosi postizanju objektivnosti istraživanja
- viši nivo pouzdanosti rezultata istraživanja
- mogućnost procene / kvantifikovanja uzoračke greške
- povećane su šanse za donošenje valjanih zaključaka o čitavoj populaciji, uopštavanjem rezultata dobijenih ispitivanjem uzorka

Mane:

- uglavnom se tiču potreba za vremenom, resursima, finansijama i ljudstvom (npr. potrebno je posedovati kompletan okvir za odabir uzorka)

1.6 Osnovni pojmovi, nastavak

Ako se na slučajan način (sa unapred određenom verovatnoćom) odabere jedna jedinica iz populacije, vrednost obeležja koju ona ima nije unapred poznata / određena. To znači da se vrednost obeležja slučajno odabrane jedinice može shvatiti kao realizacija slučajne veličine. Raspodela verovatnoća te slučajne veličina naziva se **raspodela obeležja**[5].

Statistika ('statistic') je funkcija vrednosti obeležja registrovanih na jedinicama iz odabranog uzorka, u kojoj eventualno mogu figurisati i neke poznate

[5]Zadatak matematičke statistike je određivanje raspodele obeležja ili određivanje bar nekih opštih numeričkih karakteristika te raspodele

konstante.[6]

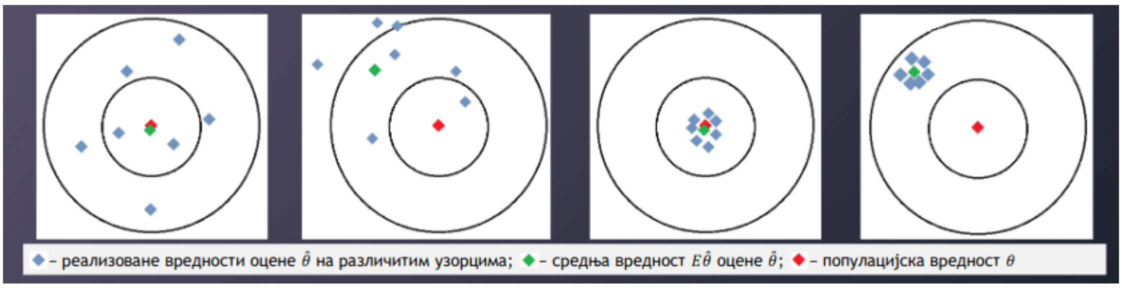
Statistike su značajne jer se često koriste za formiranje **ocena** ('estimator') parametara populacije. Realizovane vrednosti statistika su realni brojevi koji tada daju **ocene** ('estimate') nepoznatih parametara. Npr. ako je θ nepoznata populacijska vrednost onda je $\hat{\theta} = \hat{\theta(s)}$ statistika, koja predstavlja **tačkastu ocenu** ('point estimator') parametra.

Često korišćene statistike ($n(S)$ predstavlja obim uzorka S):

- uzoračka srednja vrednost
$$\overline{Y} = \frac{1}{n(S)} \sum_{k \in S} y_k$$
- uzorački total
$$T = n(S)\overline{Y}$$
- uzoračka disperzija / standardno odstupanje
$$\overline{S}^2 = \frac{1}{n(S)-1} \sum_{k \in S} (y_k - \overline{Y})^2, \overline{S} = \sqrt{\overline{S}^2}$$
- uzoračka proporcija
- uzoračka medijana, kvantili, moda

Neka je $\hat{\theta}$ tačkasta ocenapopulacijske vrednosti θ. Ona je:

- **nepristrasna** ('unbiased')
  ako jednakost $E\hat{\theta} = \theta$ važi za svaku vrednost parametra θ; ako ocena $\hat{\theta}$ nije nepristrasna onda se ona naziva **pristrasna ocena**, a vrednošću razlike $B(\hat{\theta}) := E\hat{\theta} - \theta$ meri se njena **pristrasnost**.
- **precizna** ('precise')
  ako je disperzija $D\hat{\theta} = E(\hat{\theta} - E\hat{\theta})^2$ ocene $\hat{\theta}$ mala (teži 0).
- **tačna** ('accurate')
  ako je srednje kvadratna greška $MSE(\hat{\theta}) := E(\hat{\theta} - \theta)$ ocene $\hat{\theta}$ mala.[7]



[6]Statistika je slučajna veličina sa svojom raspodelom verovatnoća, koja se naziva **uzoračka raspodela**

[7]važi i jednakost: $MSE(\hat{\theta}) = D\hat{\theta} + (B(\hat{\theta}))^2$, pa je ocena tačna ako je i precizna i nepristrasna.

2 nedelja

2.1 (Prost) slučajan uzorak

Kod (prostog) slučajnog uzorkovanja ('simple random sampling') **jedinica posmatranja = jedinica uzorkovanja**.
Neka je data populacija sa N jedinica, koje su u okviru za odabir uzorka označene brojevima iz skupa $\Omega = \{1, 2, ..., N\}$ i neka je Y obeležje od interesa. Bira se uzorak obima n.

Može biti:

- bez ponavljanja (SRSWOR)
- sa ponavljanjem (SRSWR)

2.2 SRSWOR

Predstavlja jedan od najjednostavnijih i najstarijih metoda odabira uzorka. Raspodela verovatnoća $p(\cdot)$ na kolekciji svih uzoraka $s \subset \Omega$ data je sa:

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

Dakle, ovde se svaki od $\binom{N}{n}$ mogućih podskupova skupa Ω kardinalnosti n sa podjednakom (pozitivnom) verovatnoćom može odabrati kao uzorak
Pomenuti plan obično se u praksi implementira jednim od sledeća dva ekvivalentna postupka:

- odabir uzorka vrši se kroz $n$izvlačenja („koraka“) na slučajan način, pri čemu je u svakom koraku verovatnoća izvlačenja bilo koje od jedinica, koje u ranijim koracima nisu odabrane u uzorak, ista
- odabir uzorka vrši se kroz niz **nezavisnih** izvlačenja na slučajan način **iz cele populacije**, pri čemu je u svakom koraku verovatnoća izvlačenja bilo koje od jedinica ista $\left(\frac{1}{N}\right)$,uz odbacivanje jedinica ranije odabranih u uzorak i ponavljanje koraka sve dok se ne dobije uzorak obima n

Uzorak odabran na opisani način može se prikazati i kao **uređen** niz $j_1, j_2, ..., j_n$oznaka jedinica koje su se našle u uzorku($j_k$je oznaka k-te jedinice zadržane u uzorku)

Uzorak odabran na opisani način može se prikazati i kao uređen niz $j_1, j_2, ..., j_n$ oznaka jedinica koje su se našle u uzorku (j_k je oznaka k-te) jedinice zadržane u uzorku).Pod uzorkom se, takođe, podrazumeva i pripadni niz $y_{j1}, y_{j2}, ..., y_{jn}$vrednosti posmatranog obeležja Y registrovanih na odabranim jedinicama.

Parovi $(j_k, y_{jk}), k = \overline{1, n}$, predstavljaju **podatke dobijene u istraživanju**.

2.3 SRSWR

•Odabir uzorka vrši se kroz N nezavisnih izvlačenja na slučajan način, i to uvek iz kompletne populacije, pri čemu je u svakom koraku verovatnoća

izvlačenja bilo koje od jedinica ista i jednaka $\frac{1}{N}$.
•Raspodela verovatnoća $p(\cdot)$ na kolekciji svih uzoraka $s \in \Omega^n$kao uređenih nizova dužine $n$sa dozvoljenim ponavljanjem elemenata data je sa $p(s) = N^{-n}$

2.4 Izvlačenje jedinice na slučajan način

Slučajan odabir jedinice (iz populacije u uzorak) vrši se korišćenjem **slučajnih** i **pseudoslučajnih brojeva**.
Slučajni brojevi obično se dobijaju pomoću tzv. **fizičkih generatora** (TRNG –'true random number generator').

- u makro svetu: bacanje fer novčića / kockica, slučajan izbor karte iz špila / kuglice iz kutije, rulet itd.
- u mikro svetu: prirodni fenomeni za koje važe zakonitosti kvantne mehanike, šum itd.

Oni su sadržani u tzv. **tablicama slučajnih brojeva**.

Pseudoslučajni brojevi se dobijaju pomoću tzv. **programskih generatora** (PRNG –'pseudorandom number generator'). To su računarski programi koji koriste izvestan algoritam za dobijanje niza brojeva čija svojstva, u određenoj meri, oponašaju svojstva niza slučajnih brojeva.

2.5 Novi pojmovi

- **Indikator uključenja** ('inclusion indicator')

$$I_k = \begin{cases} 1, & \text{ako je jedinica označena sa } k \text{ odabrana u uzorak} \\ 0, & \text{inače} \end{cases} \tag{2}$$

- **Verovatnoća uključenja** ('inclusion probability') prvog, odnosno drugog reda:
  π_k - verovatnoća da jedinica označena sa k bude odabrana u uzorak
  π_{kl} - verovatnoća da i jedinica označena sa k i jedinica označena sa l budu odabrane u uzorak
- **'Težina' uzorkovanja** ('sampling weight') recipročna vrednost očekivanog broja pojavljivanja jedinice označene sa k u uzorku (što se, kod uzorka bez ponavljanja, svodi na recipročnu vrednost verovatnoće uključenja prvog reda π_k) [8].

2.6 SRSWOR VS SRSWR

[8]može se interpretirati kao broj jedinica u populaciji koje reprezentuje jedinica označena sa k

SRSWOR	SRSWR
Verovatnoća uključenja prvog reda: $\pi_k = \frac{n}{N}$ za svako k	Verovatnoća uključenja prvog reda: $\pi_k = 1 - (\frac{N-1}{N})^n$ za svako k
Verovatnoća da će jedinica označena sa k biti odabrana u uzorak u j-tom koraku: $\frac{1}{N}$	Verovatnoća da će jedinica označena sa k biti odabrana u uzorak u j-tom koraku: $\frac{1}{N}$
	Verovatnoća da će jedinica označena sa k biti odabrana u uzorak više od jedanput: $1 - \left(\frac{N-1}{N}\right)^{n-1}\left(\frac{N-1-n}{N}\right)$
Očekivanibroj pojavljivanja jedinice označene sa k u uzorku: π_k	Očekivanibroj pojavljivanja jedinice označene sa k u uzorku: $\frac{n}{N}$
Verovatnoća uključenja drugog reda: $\pi_{kl} = \frac{n(n-1)}{N(N-1)}$ za $k \neq l$	Verovatnoća uključenja drugog reda: $\pi_{kl} = 1 - 2\left(\frac{N-1}{N}\right)^n + \left(\frac{N-2}{N}\right)^n$ za $k \neq l$

2.7 Pristupi prilikom zaključivanja

pristup zasnovan na metodu odabira uzorka ('design-based approach')	**pristup zasnovan na modelu ('model-based approach')**
uzoračka raspodela statistike je **diskretna raspodela verovatnoća**: ako je $\hat{\theta} = \hat{\theta}(S)$ statistika, onda važi: $P\{\hat{\theta} = m\} = \sum_{s:\hat{\theta}(s)=m} p(s)$ a njeno matematičko očekivanje i disperzija izračunavaju se po formulama: $E\hat{\theta} = \sum_m mP\{\hat{\theta} = m\} = \sum_s \hat{\theta}(s)p(s)$ $D\hat{\theta} = \sum_s (\hat{\theta}(s) - E\hat{\theta})^2 p(s)$	uzoračka raspodela statistike je **neka** jednodimenziona **raspodela verovatnoća** određena zajedničkom raspodelom verovatnoća pretpostavljenog modela populacije
nepristrasnost tačkaste ocene $E\hat{\theta}$ **u odnosu na metod odabira uzorka**	**nepristrasnost** tačkaste ocene $E\hat{\theta}$ **u odnosu na metod model**

2.8 SRSWOR VS SRSWR tačkaste ocene

	SRSWOR	SRSWR	SRSWR (u obzir se uzimaju samo različite jedinice)
tačkasta ocena $\hat{m}_Y$	$\frac{1}{n}\sum_{k\in S} y_k$	$\frac{1}{n}\sum_{k=1}^{n} y_{jk}$	$\frac{1}{n_D}\sum_k y_{(k)}$
$E\hat{m}_Y$	m_Y	m_Y	m_Y
$D\hat{m}_Y$	$\frac{\sigma^2}{n}\left(1 - \frac{n}{N}\right)$	$\frac{N-1}{N}\frac{\sigma^2}{n}$	$\sum_{k=1}^{N-1}\frac{k^{n-1}}{N^n}\sigma^2$
tačkasta ocena $D\hat{m}_Y$	$\frac{\overline{S}^2}{n}\left(1 - \frac{n}{N}\right)$	$\frac{\overline{S}^2}{n}$	

[9]

gde je σ^2 (nepoznata) populacijska disperzija, a $\overline{S}^2$ (poznata) uzoračka disperzija. [10]

2.9 Novi pojmovi

Stopa odabira uzorka, ili tzv. **razlomak uzorkovanja**('sampling fraction'), je odnos obima uzorka i obima populacije, tj. količnik $\frac{n}{N}$.

Vrednost $1 - \frac{n}{N}$ naziva se **faktor korekcije** zbog konačnosti populacije ('finite-population correction factor'). [11]

Kada su poznati matematičko očekivanje i disperzija tačkaste ocene $\hat{\theta}$ može se odrediti **koeficijent varijacije** ocene $\hat{\theta}$, definisan sa:

$$CV(\hat{\theta}) := \frac{SE(\hat{\theta})}{E\hat{\theta}}$$

i koji predstavlja meru varijabilnosti ocene.

2.10 SRSW(O)R tačaste ocene

Neka je sa S označen slučajan uzorak bez ponavljanja obima n. Kada je **pristup zasnovan na modelu**, vrlo jednostavan model populacije bio bi model u kome su slučajne veličine $Y_1, Y_2, ..., Y_N$ nezavisne i imaju istu raspodelu verovatnoća kao posmatrano obeležje Y. Ključni rezultati u vezi nepoznatom srednjom vrednošću $m_Y := EY$ obeležja Y, dati su u sledećoj tabeli:

тачкаста оцена $\widehat{m}_Y$	$\bar{Y} = \frac{1}{n}\sum_{k\in S} Y_k$
$E\widehat{m}_Y$	m_Y
$D\widehat{m}_Y$	$\frac{\sigma_Y^2}{n}$
тачкаста оцена $D\widehat{m}_Y$	$\frac{\bar{S}^2}{n}$

иста оцена може се користити за оцењивање, односно предвиђање вредности сл. величине

$$\frac{1}{N}\sum_{k=1}^{N} Y_k$$

Средње квадратна грешка предвиђања једнака је:

$$\frac{\sigma_Y^2}{n}\left(1 - \frac{n}{N}\right)$$

а њена оцена:

$$\frac{\bar{S}^2}{n}\left(1 - \frac{n}{N}\right)$$

gde je $\sigma_Y^2 := DY$ disperzija obeležja, a $\overline{S}^2$ (poznata) uzoračka disperzija.

3 nedelja

3.1 SRSW(O)R - intervalne ocene

Pretpostavlja se model populacije sa prethodnog slajda (poglavlje 2.10), pri čemu obeležje Y ima konačnu srednju vrednost i disperziju.

[9] n_D je **efektivan obim uzorka**, tj. obim redukovanog uzorka $(y_{(1)}, y_{(2)}, ..., y_{n_D})$ u kome su izostavljena eventualna ponavljanja jedinica iz originalnog uzorka

[10] može se pokazati da je $\hat{S}^2$ nepristrasna ocena σ^2

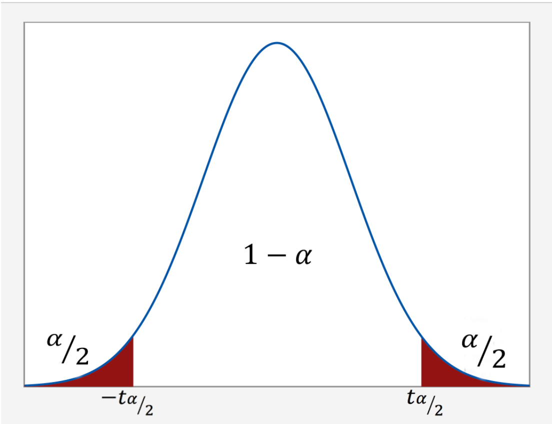
[11] U praksi se često zanemaruje kada stopa odabira uzorka ne prelazi 5%, a u mnogim slučajevima i kada je do 10%

•Ako je obim uzorka n „dovoljno veliki“ (u praksi je dovoljno već $n \geq 30$), na osnovu važenja Centralne granične teoreme, **aproksimativni** $100*(1-\alpha)\%$ (dvostrani) **interval poverenja** za nepoznatu srednju vrednost m_Y obeležja Y, dat je sa:

$$\left[\overline{Y} - z_{1-\alpha/2}\sqrt{\frac{\sigma_Y^2}{n}}, \overline{Y} + z_{1-\alpha/2}\sqrt{\frac{\sigma_Y^2}{n}}\right]$$

[12]

gde je $z_{1-\alpha/2}$ vrednost $1-\alpha/2$ - kvantila standardne normalne raspodele.



Ako je obim uzorka n manji od 30, gornja aproksimacija ne važi, pa se primenjuje egzaktan metod, koji na osnovu pretpostavki modela daje tačne intervale poverenja sa nivoom poverenja **ne manjim** od $1-\alpha$.

Specijalno, ako obeležje Y ima normalnu $\mathcal{N}(m_Y, \sigma_Y^2)$ raspodelu **tačan** $100(1-\alpha)\%$(dvostrani) **interval poverenja** za nepoznatu srednju vrednost m_Y:

- kada je σ_Y^2 poznato dat je sa:

$$\left[\overline{Y} - z_{1-\alpha/2}\sqrt{\frac{\sigma_Y^2}{n}}, \overline{Y} + z_{1-\alpha/2}\sqrt{\frac{\sigma_Y^2}{n}}\right]$$

  gde je $z_{1-\alpha/2}$ vrednost $1-\alpha/2$ - kvantila standardne normalne raspodele.

- kada je σ_Y^2 nepoznato dat je sa:

$$\left[\overline{Y} - t_{n-1;1-\alpha/2}\sqrt{\frac{\overline{S}^2}{n}}, \overline{Y} + t_{n-1;1-\alpha/2}\sqrt{\frac{\overline{S}^2}{n}}\right]$$

  gde je $t_{n-1;1-\alpha/2}$ vrednost $(1-\alpha/2)$-kvantila Studentove raspodele sa $(n-1)$ stepeni slobode.
  Za veliki obim uzorka iz obeležja sa normalnom raspodelom praktično nema razlike kada je disperzija obeležja Y poznata i kada nije, jer se tada Studentova raspodela dobro aproksimira $\mathcal{N}(0,1)$ raspodelom.

[12] σ_Y^2 ocenjujemo sa $\overline{S}^2$

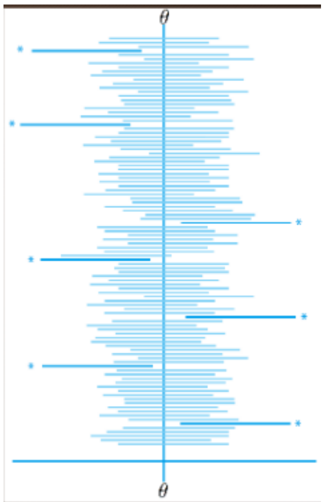
3.2 SRSWOR VS SRSWR intervalne ocene

Neka je sa $\mathcal{S}$ označen (prost) slučajan uzorak dovoljno velikog obima n. Ključni asimptotski rezultati u vezi sa intervalnom ocenom nepoznate populacijske srednje vrednosti m_Y, kada je pristup zasnovan na metodu SRSWOR, odnosno SRSWR odabira uzorka, dati su u sledećoj tabeli:

	апроксимативни $100 \cdot (1-\alpha)\%$ двострани интервал поверења	
SRSWR	$\left[\hat{m}_Y - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\bar{S}^2}{n}}, \hat{m}_Y + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\bar{S}^2}{n}}\right]$	код случајног узорка са понављањем чланови узорка су реализације независних и једнако расподељених случајних величина, па у основи лежи важење стандардне Централне граничне теореме (тј. узорачка средина која се појављује као тачкаста оцена за m_Y има приближно нормалну расподелу за довољно велико n)
SRSWOR	$\left[\hat{m}_Y - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\bar{S}^2}{n}\left(1-\frac{n}{N}\right)}, \hat{m}_Y + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\bar{S}^2}{n}\left(1-\frac{n}{N}\right)}\right]$	код случајног узорка без понављања чланови узорка су реализације случајних величина које нису независне, па се формулише специјална верзија Централне граничне теореме која се може применити у случају оваквог узорковања из коначне популације када су n, N и $N-n$ „довољно велики“; увођење појма „суперпопулације“

3.3 Interpretacija nivoa poverenja

Interpretacija nivoa poverenjaInterpretacija intervala poverenja,odnosno odgovarajućeg nivoa poverenja $1-\alpha/2$, **razlikuje se** u zavisnosti odpristupa prilikom zaključivanja.



3.4 Određivanje obima uzorka

Jedno je od prvih pitanja pri planiranju istraživanja, a odgovor na njega nije uvek jednostavan. Suštinski, radi se o odlučivanju o tome kolika je (uzoračka) greška prihvatljiva prilikom zaključivanja, pri čemu se obično mora uravnotežiti tačnost zaključivanja sa troškovima istraživanja.
Neka je $\hat{\theta}$ tačkasta ocenanepoznate populacijske vrednosti θ. Nakon preciziranja apsolutne (dozvoljene) greške('margin of error') Δ za zadati nivo poverenja $1-\alpha$, pitanje se svodi na određivanje vrednosti n tako da važi

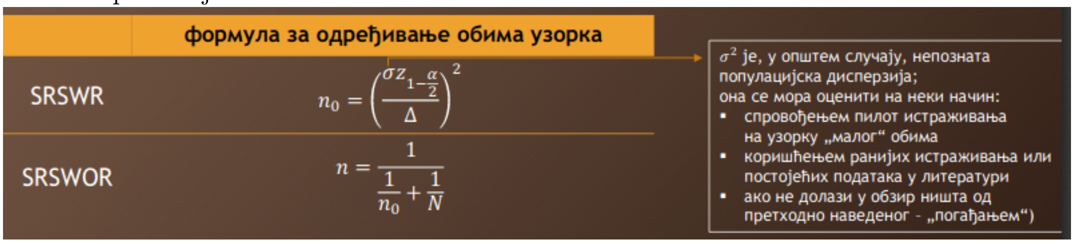
$$P\{|\hat{\theta}-\theta| > \Delta\} < \alpha$$

Npr. ako je $\hat{\theta}$ nepristrasna, normalno raspodeljena ocena parametra θ onda

$$P\left\{\frac{|\hat{\theta}-\theta|}{\sqrt{D\hat{\theta}}} > z_{1-\frac{\alpha}{2}}\right\} = P\left\{|\hat{\theta}-\theta| > z_{1-\frac{\alpha}{2}}\sqrt{D\hat{\theta}}\right\} = \alpha$$

pa kako disperzija ocene $\hat{\theta}$ opada sa obimom uzorka n, onda će gornja nejednakost biti zadovoljena ako se odabere dovoljno veliko n tako da važi

Najjednostavnija jednačina za određivanje obima uzorka za ocenjivanje nepoznate populacijske srednje vrednosti m_Y, tako da se postigne apsolutna greška ne veća od Δ sa poverenjem $1 - \alpha$, može se dobiti na osnovu aproksimativnih intervala poverenja:



	формула за одређивање обима узорка
SRSWR	$n_0 = \left(\frac{\sigma z_{1-\frac{\alpha}{2}}}{\Delta}\right)^2$
SRSWOR	$n = \frac{1}{\frac{1}{n_0} + \frac{1}{N}}$

σ^2 je, у општем случају, непозната популацијска дисперзија;
она се мора оценити на неки начин:
- спровођењем пилот истраживања на узорку „малог“ обима
- коришћењем ранијих истраживања или постојећих података у литератури
- ако не долази у обзир ништа од претходно наведеног - „погађањем“)

Pored opisanog kriterijuma određivanja obima uzorka zadavanjem apsolutne greške ocene, postoje i drugi kriterijumi i to:

- zadavanjem širine intervala poverenja
- zadavanjem gornje granice disperzije / standardne greške ocene
- zadavanjem relativne greške ocene

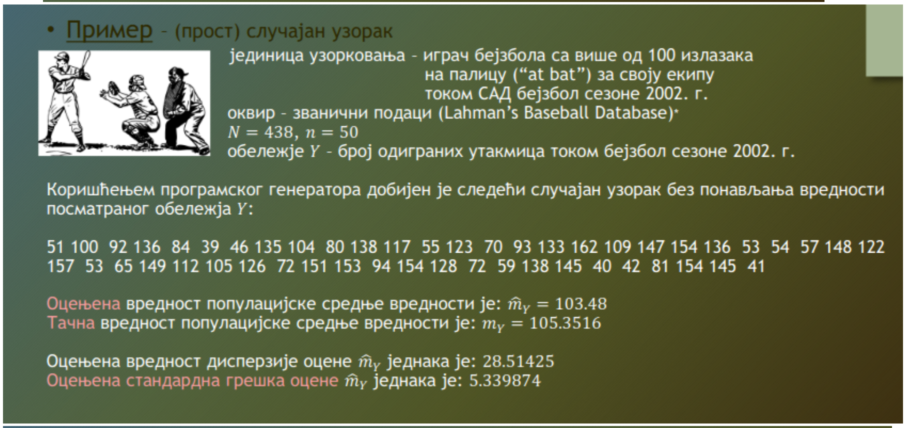
Neka je $\hat{\theta}$ tačkasta ocenane poznate populacijske vrednosti θ. Nakon preciziranja **relativne greške** p za zadati nivo poverenja $1 - \alpha$, pitanje se se svodi na određivanje vrednosti n tako da važi

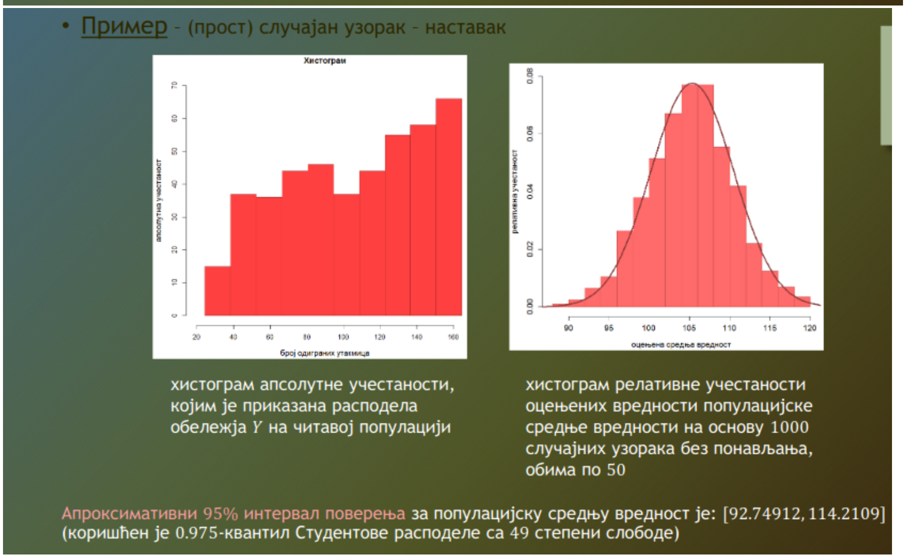
$$P\left\{\frac{|\hat{\theta} - \theta|}{\theta} > p\right\} < \alpha$$

- zadavanjem koeficijenta varijacije ocene
- zadavanjem troškova uzorkovanja

Rezultati koji se tiču nepoznatog populacijskog totala τ_Y potpunosti su analogni prikazanim rezultatima u vezi sa nepoznatom populacijskom srednjom vrednošću m_Y.

4 nedelja (primeri)





4.1 SRSW(O)R ocenjivanje proporcije

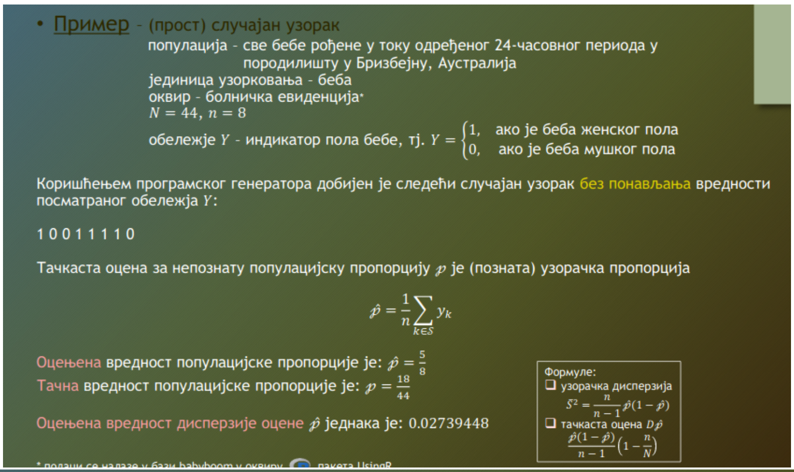
Pretpostavi se da je istraživač zainteresovan za neko kvalitativno svojstvo jedinica posmatrane populacije.
Populacijska proporcija p predstavlja udeo jedinica populacije koje pripadaju posmatranom nivou kategoričke promenljive od interesa, tj.

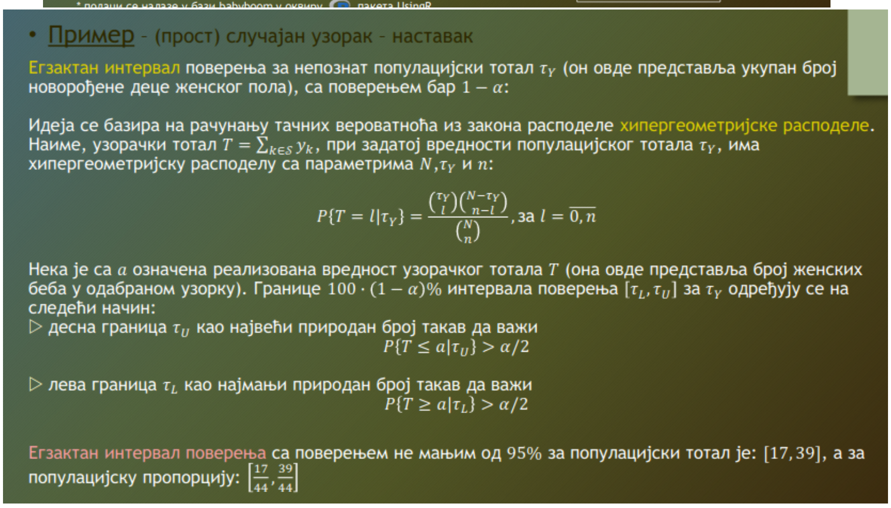
$$p = \frac{\text{broj jedinica u populaciji koje poseduju dato svojstvo}}{N}$$

Sa statističkog stanovišta cilj je oceniti nepoznati parametar p, pri čemu je obeležje Y indikatorska slučajna veličina, data sa:

$$Y = \begin{cases} 1, & \text{ako jedinica poseduje dato svojstvo} \\ 0, & \text{inače} \end{cases} \tag{3}$$

Za ovako definisano obeležje od interesa, populacijska srednja vrednost zapravo je populacijska proporcija, pa se mogu koristiti svi ključni rezultati u vezi sa ocenjivanjem nepoznate srednje vrednosti.





- **Пример** - (прост) случајан узорак - наставак

Сада се претпостави да је, коришћењем програмског генератора, добијен следећи случајан узорак обима $n = 8$ са понављањем вредности посматраног обележја Y:

1 0 0 0 1 0 1 0

Егзактан интервал поверења за непознату популацијску пропорцију p, са поверењем бар $1 - \alpha$:

Идеја се базира на рачунању тачних вероватноћа из закона расподеле биномне расподеле. Наиме, узорачки тотал $T = \sum_{k \in s} y_k$, при задатој вредности популацијске пропорције p, има биномну $\mathcal{B}(n, p)$ расподелу.

Нека је са a означена реализована вредност узорачког тотала T. Ако је $a \neq 0$ и $a \neq n$, границе $100 \cdot (1 - \alpha)\%$ интервала поверења (p_L, p_U) за p одређују се на следећи начин:

▷ десна граница p_U као решење једначине

$$\sum_{l=0}^{a} \binom{n}{l} p_U^l (1 - p_U)^{n-l} = \alpha/2$$

▷ лева граница p_L као решење једначине

$$\sum_{l=0}^{n-a} \binom{n}{l} (1 - p_L)^l p_L^{n-l} = \alpha/2$$

Егзактан интервал поверења са поверењем не мањим од 95% за популацијску пропорцију је: [0.08523341,0.7551368]

- **Пример** - (прост) случајан узорак

Планирано је истраживање јавног здравља на широј територији једног великог града ради процене удела деце, узраста 0-14 година, на којој није спроведена имунизација (вакцинација) против дечје парализе. Организатори овог пројекта желели би да постигну да узорачки удео неадекватно имунизоване деце са вероватноћом бар 98% одступа, по апсолутној вредности, од тачног удела p за највише 0.05. Одредити колики би требало да буде обим узорка.

С обзиром на претпоставке и овде се може изоставити фактор корекције и применити нормална апроксимација

Формула за одређивање траженог обима узорка своди се на:

$$n_0 = \frac{p(1 - p) z_{1-\frac{\alpha}{2}}^2}{\Delta^2},$$

где је апсолутна грешка $\Delta = 0.05$.

Приметити да наведена формула зависи од непознате популацијске пропорције p. Најгрубља могућа процена за p је $p = 0.5$, па је тада најмањи обим узорка, за задату тачност, дат вредношћу израза:

$$\frac{z_{1-\frac{\alpha}{2}}^2}{4\Delta^2}$$

Обим узорка за оцењивање непознате популацијске пропорције је: 542

- **Пример** - (прост) случајан узорак

У децембру 2018. г, на националном нивоу, испитиван је став јавног мњења о увођењу стандарда за аутомобиле са циљем повећања енергетске ефикасности горива.
У те сврхе спроведена је анкета на (простом) случајном узорку обима $n = 1012$ одраслих становника који поседују возачку дозволу. Њих $a = 810$ изјаснило се у корист увођења оваквих стандарда.

Нека је p непозната популацијска пропорција возача који подржавају увођење стандарда (она се може схватити и као (непозната) вероватноћа да (произвољан) возач подржава увођење стандарда).

С обзиром на то да је истраживање вршено на националном нивоу, те да је, према томе, обим узорка занемарљиво мали у односу на обим популације, може се свуда изоставити фактор корекције $1 - \frac{n}{N}$

Одређује се апроксимативни 95% интервал поверења за популацијску пропорцију. То се врши на основу важења Централне граничне теореме, јер је n довољно велико:

$$P\left\{\frac{|\hat{p} - p|}{\sqrt{p(1 - p)}} \cdot \sqrt{n} \leq z_{1-\frac{\alpha}{2}}\right\} \approx 1 - \frac{\alpha}{2}$$

Овде важи: $D\hat{p} \approx \frac{p(1 - p)}{n}$

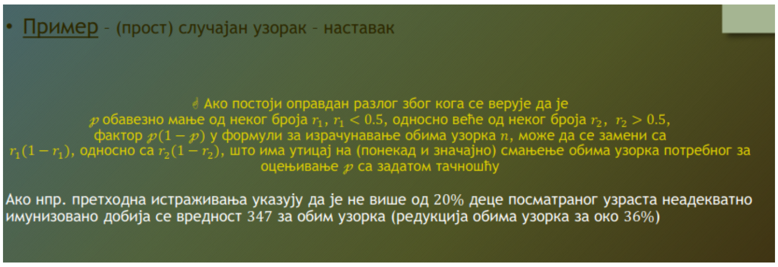
Облик интервала поверења је $[p_L, p_U]$, где су p_L и p_U решења квадратне једначине:

$$\left(n + z_{1-\frac{\alpha}{2}}^2\right) p^2 - \left(2n\hat{p} + z_{1-\frac{\alpha}{2}}^2\right) p + n\hat{p}^2 = 0$$

по промењљивој p

Оцењена вредност популацијске пропорције је: $\hat{p} = \frac{810}{1012}$

Апроксимативни 95% интервал поверења за популацијску пропорцију је: $[0.775, 0.824]$



4.2 SRSWOR - ocenjivanje parametara populacije

Neka je obim posmatrane populacije N i neka je $N_l$broj jedinica te populacije koje pripadaju njenoj l-toj potpopulaciji. Neka je sa y_{lk} označena vrednost obeležja Y entiteta označenog sa k, koji potiče iz l-te potpopulacije. Potpopulacijski total, odnosno srednja vrednost,dati su sa

$$\tau_l = \sum_{k=1}^{N_l} y_{lk}, \quad m_l = \frac{1}{N_l}\sum_{k=1}^{N_l} y_{lk} \tag{4}$$

tačkasta ocena	$\frac{1}{n_l}\sum_{k\in S_l} y_{lk}$
$E\hat{m}_l$	m_l
$D\hat{m}_l$	$\sigma_l^2\left(E\left(\frac{1}{n_l}\right)-\frac{1}{N_l}\right)$
tačkasta ocena $D\hat{m}_l$	$\frac{\overline{S}_l^2}{n_l}\left(1-\frac{n_l}{N_l}\right)$

kada je obim $N_l l$-te potpopulacije nepoznat(što je često slučaj),faktor korekcije zbog konačnostipopulacije $1-\frac{n_l}{N_l}$, zamenjuje se svojimmatematičkim očekivanjem $1-\frac{n}{N}$.

σ_l^2 (nepoznata) potpopulacijska disperzijaza jedinice iz l-te potpopulacije,a S_l^2 (poznata) uzoračka disperzijaza jedinice iz poduzorka S_l.

Ako je **poznat** obim N_l l-te potpopulacije rezultati koji se tiču nepoznatog potpopulacijskog totala lako se dobijajuiz onih za potpopulacijsku srednju vrednost m_l korišćenjem veze: $\tau_l = N_l m_l$.
Ako **nije poznat** obim N_l kao nepristrasna ocena za τ_l može se koristiti statistika:

$$\hat{\tau}_l = \frac{N}{n}\sum_{k\in S_l} y_{lk}$$