Skripta iz Uvoda u teoriju uzoraka

22. februar 2020

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$-N \to +\infty$

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

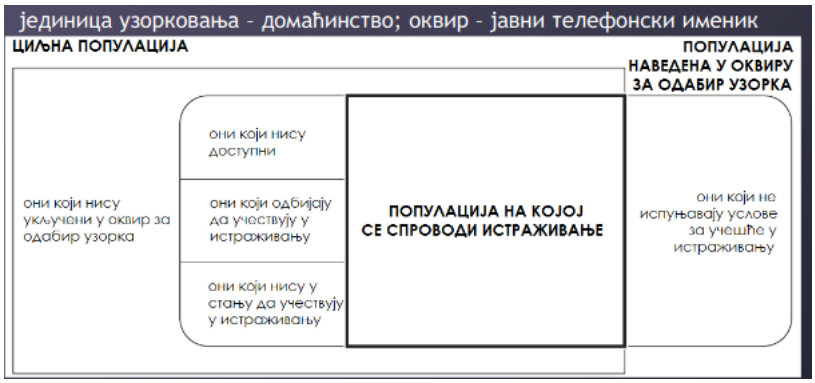
Obim uzorka ('samplesize') n [2]

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

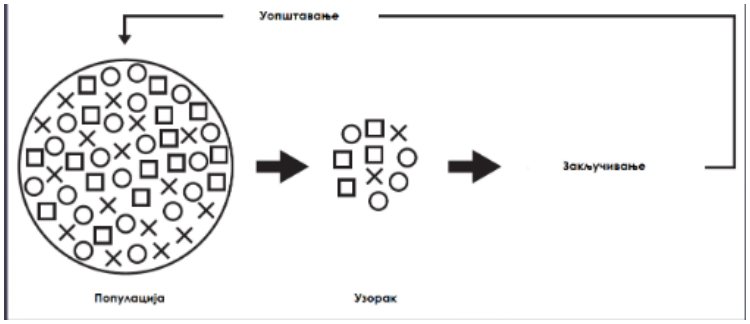
[1] Nadalje se pretpostavlja: target population = study population i $N < +\infty$

[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

mogućnost potpunog ispitivanja populacije istraživač se obično opredeljuje za **delimično** ispitivanje (proučavanje uzorka) jer je (u odnosu na potpuno ispitivanje):

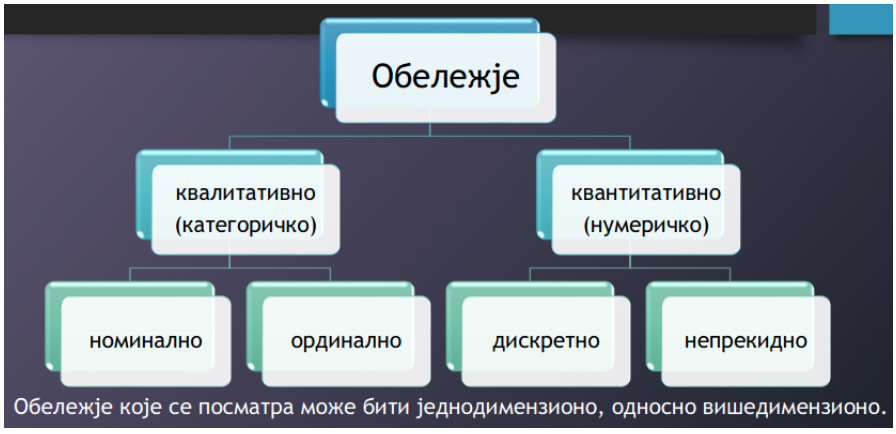
- jeftinije
- brže
- kontrola tačnosti prikuplenih podataka je jednostavnija i lakša



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

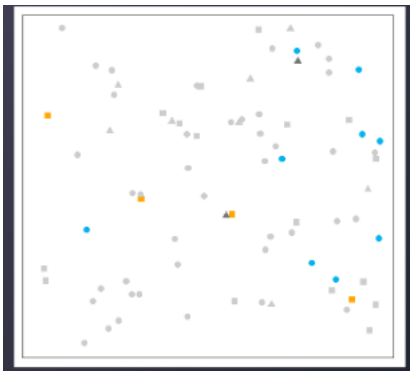
$$\sigma_Y = \sigma = \sqrt{\sigma_Y^2}$$

- 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

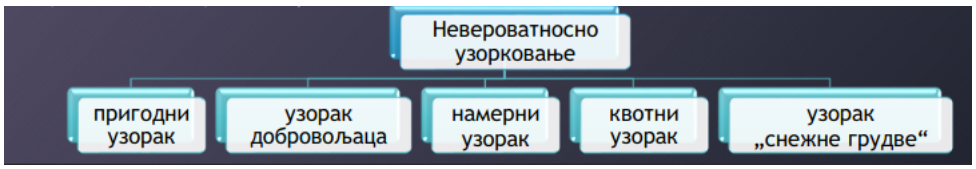
Metod odabira uzorka je postupak kojim se biraju elementi populacije u uzorak, uz određivanje adekvatnog obima uzorka.
Ovi metodi se mogu podeliti u dve grupe:

- **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 $\hat{\theta} = \theta\hat{(}s)$ 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

$$T = n(S)\overline{Y}$$

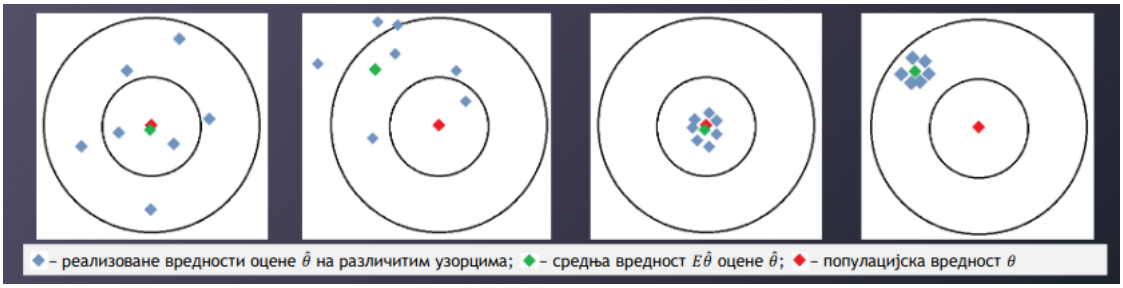
- uzoračka disperzija / standardno odstupanje

$$\overline{S}^2 = \frac{1}{n(S)-1} \sum_{k \in S} (y_k - \overline{Y})^2, \overline{S} = \sqrt{\overline{S}^2}$$

- 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.