

Skripta iz Uvoda u teoriju uzoraka

24. 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 objekt 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')

Uzorak ('sample') - podskup populacije; sadrži izvesne entitete koji potiču iz populacije, na bazi čijeg proučavanja će se izvoditi zaključci o čitavoj populaciji

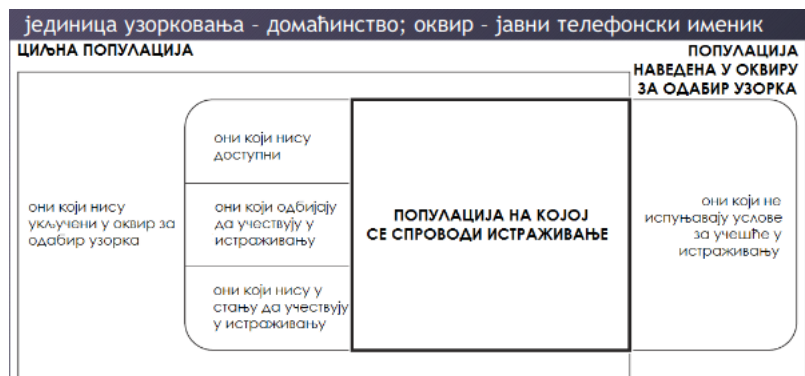
Obim uzorka ('samplesize') n ²

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

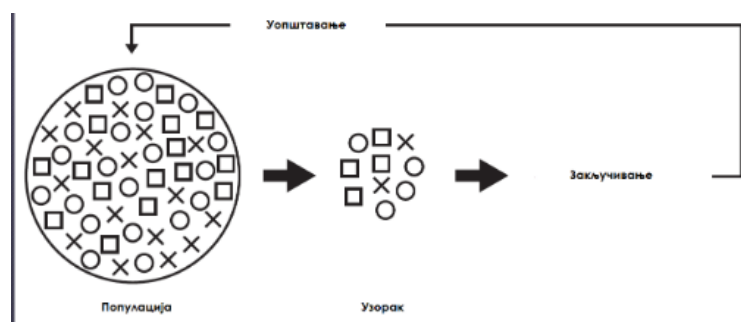
¹Nadalje se pretpostavlja: target population = study population i $N < +\infty$

²Uvek konačna vrednost

³U opštem slučaju nije isto što i jedinica posmatranja, koja predstavlja osnovni objekat posmatranja i prikupljanja informacija. Jedinice uzorkovanja su međusobno disjunktne skupovi entiteta

могућност потпуног испитивања популације истраживач се обично определjuje за **делимично** испитивање (пroučавање узорка) jer je (у односу на потпуно испитивање):

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

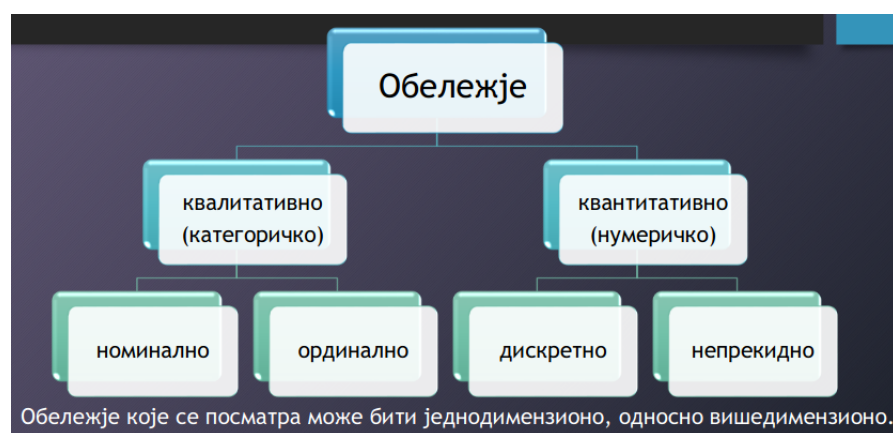


Termin populacija odnosi se na skup entiteta istovrsnih u odnosu na jedno ili više zajedničkih svojstava, koja se mogu posmatrati. Ipak, entiteti, iako **istovrsni, nisu istovetni**.

Одређивање популације представља значајну и, neretko, tešku fazu истраживања. Популација мора бити дефинисана: појмовно (у смислу свог садржаја - шта су ентитети, а шта јединице узorkовања?), просторно и временски.

Оbeležje ('study variable') - посматрана заједничка карактеристика свих ентитета у популацији, тј. прецизније, извесно варијабилно својство од интереса, које је одређено за сваки ентитет у популацији.⁴

1.3 Tipovi obeležja



⁴Obeležje 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
 - * koncentracija 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 proizvođača.
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 odabir uzorka označene brojevima iz skupa $\omega = 1, 2, \dots, N$ (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

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

, 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

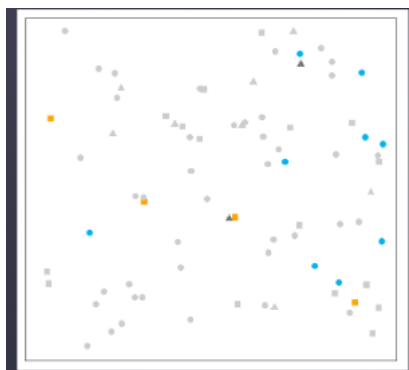
- populacijska proporcija ('population proportion')
- populacijska medijana, kvantili, moda...

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 „uvećanu“ sliku jednog dela populacije. Uzorak sa ovom osobinom verno odslikava strukturu populacije koju predstavlja, „izgledajući“ kao i populacija u svim aspektima relevantnim 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

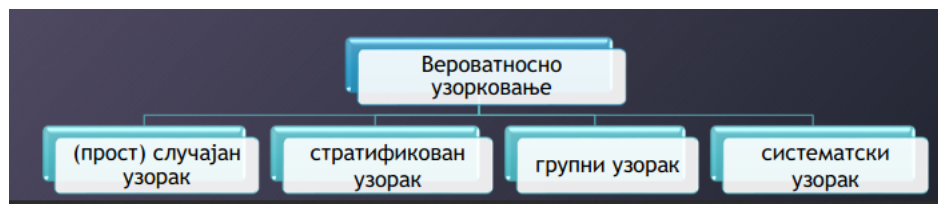
Ovakvi metodi uzorkovanja zasnivaju se na teoriji verovatnoća, tj. na „planiranoj“ slučajnosti. Mogući uzorci su faktički matematički konstruisani, i za svakog od njih poznata je verovatnoća da bude odabran. Dakle, uzorkovanje se vrši u skladu sa raspodelom verovatnoća, definisanom na kolekciji svih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

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

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

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

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



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 veličine 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

⁵Zadatak 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

$$\bar{Y} = \frac{1}{n(S)} \sum_{k \in S} y_k$$

- uzorački total

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

- uzoračka disperzija / standardno odstupanje

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

- uzoračka proporcija
- uzoračka medijana, kvantili, moda

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

- **nepistrasna** ('unbiased')
ako jednakost $E\hat{\theta} = \theta$ važi za svaku vrednost parametra θ ; ako ocena $\hat{\theta}$ nije nepistrasna 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 $MSE(\hat{\theta}) := E(\hat{\theta} - \theta)^2$ ocene $\hat{\theta}$ mala.⁷



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

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

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 $\Omega = \{1, 2, \dots, N\}$ 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:

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

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 nizvlačenja („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, \dots, 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, \dots, 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 $y_{j_1}, y_{j_2}, \dots, y_{j_n}$ vrednosti posmatranog obeležja Y registrovanih na odabranim jedinicama.

Parovi $(j_k, y_{j_k}), k = \overline{1, n}$, 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')

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

- **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

⁸može 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: $\pi_k = 1 - \left(\frac{N-1}{N}\right)^n$ 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’)
<p>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: $E\hat{\theta} = \sum_m m P\{\hat{\theta} = m\} = \sum_s \hat{\theta}(s) p(s)$ $D\hat{\theta} = \sum_s (\hat{\theta}(s) - E\hat{\theta})^2 p(s)$</p>	<p>uzoračka raspodela statistike je neka jednodimenziona raspodela verovatnoća određena zajedničkom raspodelom verovatnoća pretpostavljenog modela populacije</p>
nepriistrasnost tačkaste ocene $E\hat{\theta}$ u odnosu na metod odabira uzorka	nepriistrasnost 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 $D\hat{m}_Y$	$\frac{\bar{S}^2}{n} \left(1 - \frac{n}{N}\right)$	$\frac{\bar{S}^2}{n}$	

9

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

2.9 Novi pojmovi

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

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