Uzorak

U statističkom smislu populacija ili osnovni skup obuhvata sve statističke jedinice koje ispoljavaju obeležje koje je predmet istraživanja. Populacija može biti konačna i beskonačna. Definisanje populacije postiže se popisom statističkih jedinica. Međutim, tokom istraživanja često ne postoje mogućnosti da se obuhvate sve statističke jedinice, a ako i postoje, radi se o zamornom, skupom i dugotrajnom procesu. Kada se radi o beskonačnoj populaciji ili prebrojivoj, ali vrlo velikoj populaciji nepostojanje popisa se savladava višefaznim uzorkovanjem.

Dakle, u medicinskim istraživanjima se najčešće koriste uzorci koji reprezentuju osnovni skup. Primena uzorka zasniva se na teoriji verovatnoće i predstavlja vrstu delimičnog prikupljanja podataka.

Populacija (osnovni skup, univerzum) predstavlja skup svih jedinica (elemenata, članova) s određenim zajedničkim karakteristikama. Jedinice posmatranja u populaciji nazivaju se statističke jedinice ili entiteti. Razlikujemo ih prema njihovim obeležjima, koje još nazivamo i atributi.

Uzorak je deo osnovnog skupa na osnovu čijeg proučavanja donosimo zaključke o celom osnovnom skupu, tj. deo osnovnog skupa na osnovu kojeg se procenjuju parametri osnovnog skupa (parametri: $\overline{X}$ *, SD,* SD^2*...).*

Adekvatan uzorak mora da ispuni principe nepristrasnosti, reprezentativnosti i ekonomičnosti.

Nepristrasnost podrazumeva podjedanku verovatnoću da svaki od elemanata uđe u uzorak. Ona se postiže načinom i metodom odabiranja uzorka koji su razrađeni u statističkoj praksi, a koji se baziraju na teoriji verovatnoće i postavkama slučajnog kombinovanja elemenata.

Reprezentativnost podrazumeva da uzorak treba da obuhvati one statističke jedinice koje će u sebi nositi sve karakteristike osnovnog skupa, odnosno one jedinice čija obeležja, kada se izbroje ili izmere i iz njihovih vrednosti izračunaju odgovarajući parametri (aritmetička sredina, standardna devijacija i dr.), budu ista ili približno ista kao i pravi parametri osnovnog skupa.

Reprezentativnost uzorka se postiže adekvatnom veličinom uzorka i objektivnim načinom izbora jedinica uzorka.

Određivanje adekvatne veličine uzorka je jedan od najznačajnijih zadataka pri dizajnu istraživanja koja može bitno da utiče na donošenje preciznih zaključaka. Veličina uzorka zavisi od tri grupe faktora:

- Nivoa pouzdanosti
- Varijabilnosti osnovnog skupa
- Maksimalno dozvoljene greške

1. Nivo pozdanosti – U praksi se najčešće primenjuje nivo pouzdanosti p=0,95(95%) i p=0,99(99%), kome odgovara standardizovana z-vrednost od ±1,96 i ±2,58 (zbog jednostavnosti često se uzimaju vrednosti za z od 2 i 3). Ukoliko želimo da nivo pouzdanosti bude veći potrebno je da i uzorak bude veći, i obrnuto.

2. Varijabilnost osnovnog skupa – Varijabilnost osnovnog skupa procenjuje se na osnovu SD uzorka iz koje se izračunava standardna greška osnovnog skupa. Što je veća varijabilnost obeležja, to je skup manje homogen, pa je za odgovarajuću preciznost i pouzdanost procene potreban veći uzorak.

3. Maksimalno dozvoljena greška određena je centrom serije i tačkom najvećeg odstupanja od nje u jednom pravcu. Odnosno, to je maksimalan iznos greške koji je istraživač spreman da toleriše. Mala dozvoljena greška zahteva veći uzorak i obrnuto.

Ekonomičnost je princip koji nameću finansijska i vremenska ograničenja. Veliki uzorak zahteva više finansijskih i ljudskih resursa, kao i vremena za ispitivanje. Princip ekonomičnosti je suštinski suprotan principu reprezentativnosti.

Pravilan način izbora uzorka

Pravilan način izbora uzorka isključuje subjektivnost. Subjektivnost se anulira metodama slučajnog odabiranja jedinica u uzorak. Slučajnost znači da treba primeniti takav način odabiranja koji isključuje pristrasnost i pri kome svaka jedinica osnovnog skupa ima podjednaku verovatnoću da uđe u uzorak, odnosno svaki uzorak ima istu šansu da bude izabran kao predstavnik skupa.

Prema tehnici uzorkovanja, uzorci se mogu podeliti na:

1. Prost – slučajan uzorak
2. Sistematski uzorak
3. Stratifikovani uzorak

Prost – slučajan uzorak se formira metodom tablica slučajnih brojeva i metodom lutrije. Najpoznatije tablice slučajnih brojeva su Fisherova, Tippetova, Kendallova i dr. Tablice slučajnih brojeva sadrže niz brojeva koji su na slučajan način raspoređeni u kolone i redove. Kako bi se formirao uzorak na osnovu tablica slučajnih brojeva potrebno je prvo napraviti spisak svih statisitčkih jedinica osnovnog skupa, takozvani «okvir skupa», a zatim numerisati statističke jedinice od 1 do N. Tada se, na slučajan način odredi prvi broj u tablici koji predstavlja i prvu jedinicu u uzorku, a zatim se redom izdvaja n brojeva manjih od N.

Jedan deo Tippetove tablice slučajnih brojeva:

56**24**	4167	9524	15**45**	1396	72**03**	5356
1300	2693	27**30**	7483	3408	2762	3563
1089	69**13**	7691	0560	52**46**	11**12**	51**07**
6008	81**26**	42**33**	8776	2754	91**43**	14**05**

Uzmimo na primer da iz skupa od 50 bolesnika treba formirati uzorak od 10 ispitanika (20%), pod pretpostavkom da smo slučajno izabrali kao prvi broj 5624, zatim krećemo udesno za izborom jedinica za uzorak. Biramo dvocifrene brojeve manje od 50, i to sa poslednje dve cifre iz grupe. Na taj način biramo sledeće jedinice:

24 45 3 30 13 46 12 7 26 33

Kako su se brojevi 24 i 8 ponovili, njih preskačemo, i nastavljamo biranje dva sledeća broja. Na osnovu ove tablice izabrani su redni brojevi bolesnika koji će ući u ispitivanje.

Metoda lutrije je nešto jednostavnija za odabiranje slučajnog uzorka, ali ne ispunjava potpunu zakonitost slučajnosti. Za ovu metodu takođe se formira okvir skupa, a zatim se na karticama ispišu redni brojevi, nakon mešanja svih kartica, vrši se izvlačenje. Kartice sa izvučenim redni brojevima određuju koji će bolesnici ući u ispitivanje. Za naš prethodni primer potrebno je izvući 10 kartica.

Na ovaj način omogućava se da sve statističke jedinice imaju jednaku verovatnoću, odnosno jednaku šansu da postanu deo uzorka, čime se obezbeđuje reprezentativnost uzorka.

Prost – slučajan uzorak predstavlja jednostavan, korektan i relativno lak način uzorkovanja. Nedostatak ovog načina formiranja uzorka ispoljava se kada osnovni skup ima veliki broj jedinica, tada je otežana identifikacija onih jedinica koje ulaze u uzorak.

Prednost u takvoj situaciji ima **sistematski uzorak**. Sistematski uzorak se određuje po nekom unapred **određenom sistemu.** Formiranje sistematskog uzorka je višefazni proces. Prvo se sastavlja spisak statističkih jedinica (okvir skupa), zatim se određuje stopa ili korak.

Stopa se određuje po formuli:

$R = \frac{N}{n}$, R= stopa (korak),

N= ukupan br.statističkih jedinica u osnovnom skupu,

n=br.statističkih jedinica u uzorku

Nakon određivanja stope, vrši se određivanje slučajnog početka. Finalni korak je izvlačenje svake R-te jedinice i formiranje uzorka. Sistematski uzorak je posebno pogodan za korišćenje kada su statističke jedinice poređane po veličini.

Primena **stratifikovanog uzorka** je karakteristična za izrazito heterogene skupove. Formiranjem stratuma dobijaju se homogeni podskupovi iz kojih se zatim biraju jedinice koje obrazuju prost slučajan uzorak. Stratumi se određuju prema nekom određenom kriterijumu, time se postiže homogenost i između stratuma se pokazuju prave varijacije.

$n_s = n \times \frac{N_S}{N}$ n – br. statističkih jedinica u uzorku stratuma

ns – ukupan br.statističkih jedinica u uzorku

Ns – ukupan br. statističkih jedinica u stratumu

N – ukupan br. jedinica svih strauma

Stratifikacija osnovnog skupa omogućava formiranje reprezentativnog uzorka, analizu svih delova skupa, smanjuje se slučajna greška izbora jer se izbor vrši iz homogenih skupova. Izazov prilikom obrazovanja stratifikovanog uzorka predstavlja odabir dovoljnog broja stratuma i sam način stratifikacije. Treba težiti da se skup podeli na što više homogenih stratuma, jer su na taj nain i varijanse ovih podskupova male, samim tim mala je greška ocene aritmetičke sredine i drugih parametara osnovnog skupa. U svakodnevnom radu, stratifikacija se vrši u skladu sa konkretnim problemom uz prethodno proučavanje pojave i njene strukture.

Prema veličini uzoraci se dele na:

- Mali uzorke n < 30 jedinica
- Veliki uzorke n > 30 jedinica

Zavisni i nezavisni uzorci

U zavisnosti od koncepta istraživanja dešava se da se neko obeležje ispituje jednom ili više puta na jedinicama posmatranja. U zavisnosti od toga koliko se puta neko obeležje registruje na jedinicama posmatranja uzorci se dele na zavisne i nezavisne.

Kod zavsinih uzoraka jedno obeležje se registruje dva ili više puta na istim statističkim jedinicama. To praktično znači da su statističke jedinice u uzorku u jednom trenutku kontrolna, a u drugom eksperimentalna grupa. Primer za to je ispitivanje efikasnosti neke terapije. Npr. određuje se glikemija kod dijabetičara pre davanja nove antidijabetesne terapije, a zatim se kod istih osoba određuje glikemija nakon delovanja terapije. Na taj način može uspešno da se prati efekat terapije.

Kod nezavisnih uzoraka jedno obeležje se registruje samo jedanput na jedinicama posmatranja. Na primer određujemo vrednost holesterola kod ispitanika u selu i u gradu. Jedan uzorak čine ispitanici koji žive na selu, a drugi ispitanici koji su iz grada. Zaključivanje na osnovu zavisnog uzorka ima veći pouzdanost nego zaključivanje na osnovu nezavisnog uzorka, jer se pri tome vrši bolja kontrola svih individualnih pridruženih faktora, koji se za vreme istraživanja ne menjaju ili su promene neznatne. Zbog toga je i potrebna veličina zavisnih uzoraka manja nego veličina nezavisnih uzoraka.

Ocena aritmetičke sredine osnovnog skupa na osnovu uzorka

Primena uzorka ima za cilj da pruži informacije o osnovnom skupu. Kako bi se dobile precizne i validne informacije potrebno je da osobine uzorka bude što sličnije osobinama osnovnog skupa.

Iz jednog osnovnog skupa sa ***n*** statističkih jedinica moguće je formirati ***k*** jednakih uzoraka. Parametri tih uzoraka međusobno variraju formirajući distribuciju mogućih vrednosti tog parametra (sampling distribucija).

Osnovni parametri su aritmetička sredina, medijana, varijansa i standardna devijacija.

Uzmimo primer da neki osnovni skup čine telesne mase 16 novorođenčeta u nekom porodilištu.

2,75 3,05 3,30 3,35 4,15 3,70 2,95 3,40

3,95 3,15 3,85 3,55 3,80 4,05 2,65 3,15

Prosečna telesna masa novorođenčadi je $\overline{X}$ =3,42, a SD=0,46.

Ako sada formiramo 4 uzorka po 4 člana i odredimo njihove srednje vrednosti i standardne devijacije dobićemo sledeće.

		$\overline{X}$	SD
I uzorak	2,75 3,05 3,30 3,35	3,11	0,27
II uzorak	4,15 3,70 2,95 3,40	3,55	0,44
III uzorak	3,95 3,15 3,85 3,55	3,62	0,36
IV uzorak	3,80 4,05 2,65 3,15	3,41	0,56

Prosečne vrednosti uzoraka kreću se od 3,11 – 3,62.

Ako bismo sada, formirali dva uzorka sa po 8 jedinica, srednja vrednost prvog uzorka je $\overline{X}$=3,33, a za drugi uzorak $\overline{X}$=3,52. U ovom slučaju aritmetičke sredine su bliže aritmetičkoj sredini osnovnog skupa. Što pokazuje da se dobijaju sve približnije vrednosti što je uzorak veći.

Kada, saberemo aritmetičke sredine sva četri uzorka:

$$\overline{\overline{x}} = \frac{3{,}11 + 3{,}55 + 3{,}62 + 3{,}41}{4} = 3{,}42$$

i saberemo aritmetičke sredine dva uzorka od po 8 jedinica:

$$\overline{\overline{x}} = \frac{3{,}33 + 3{,}52}{2} = 3{,}42$$

dobijamo istu aritmetičku sredinu.

Što praktično pokazuje da je aritmetička sredina svih aritmetičkih sredina uzoraka jednaka aritmetičkoj sredini osnovnog skupa (prvo pravilo).

$$\overline{\overline{x}} = \overline{x}_{\text{osnovnog skupa}}$$

Aritmetičke sredine uzoraka raspoređuju se oko aritmetičke sredine osnovnog skupa u vidu normalne distribucije.

Raspored nije uslovljen unutrašnjim rasporedom jedinica osnovnog skupa, mada se mogu izdvojiti nekoliko oblika normalnosti:

- Ako je raspored jedinica unutar osnovnog skupa normalan, raspored aritmetičkih sredina uzoraka oko aritmetičke sredine osnovnog skupa ima sve osobine Gausovog normalnog rasporeda, bez obzira na veličinu uzorka.

- Iste osobine pokazuje i distribucija aritmetičkih sredina jednakih uzoraka dobijenih iz osnovnog skupa koji ne pokazuje normalan raspored, ako su uzorci veliki. Obično ovaj uslov zadovoljavaju uzorci veći od 30 jedinica (n>30).

- Aritmetičke sredine jednakih malih uzoraka (n<30) dobijenih iz skupa koji ne pokazuje normalan raspored, takođe čine jedan oblik normalnog rasporeda oko aritmetičke sredine osnovnog skupa, poznat pod nazivom Student-ov t raspored.

Na osnovu praktično pokazanih odnosa aritmetičkih sredina uzoraka i aritmetičkih sredina osonovnog skupa definisana je centralna granična teorema koja glasi:

Bez obzira na oblik rasporeda osnovnog skupa raspored aritmetičkih sredina uzoraka se približava normalnom rasporedu kako broj i veličina uzoraka raste.

Varijansa osnovnog skupa $SD^2_{os}=0,21$,

Varijansa distribucije frekvencije aritmetičkih sredina uzoraka je izračunata po formuli:

$$SD^2_{uzorka}=\frac{\sum f\overline{X^2}}{\text{broj uzoraka}}-\overline{\overline{X}}^2\text{, } SD^2_{uzorka}=0,017,$$

$SD_{uzorka}=0,13$.

Vidi se iz primera da je varijansa osnovnog skupa veća od varijanse distibucije frekvencije aritmetičkih sredina, jer su uzorci kompenzovali varijabilitet kroz svoje aritmetičke sredine.

Varijansa osnovnog skupa je veća od varijanse distribucije aritmetičkih sredina onoliko puta, kolika je veličina uzoraka (drugo pravilo).

$$SD^2_{os}=n\cdot SD^2_{uz}$$

Varijanse uzoraka čine takvu distibuciju oko prave varijanse osnovnog skupa, da njihova aritmetička sredina odgovara varijansi skupa.

$$\overline{X}_{SD^2_{uzorka}}=\frac{\sum f\cdot SD^2_{uz}}{\sum f}=SD^2_{os}$$

Standardna greška

Iz osobina distribucije aritmetičkih sredina izvedena je formula standardne devijacije distribucije aritmetičkih sredina, ovaj parametar naziva se standardna greška.

$$SG = \frac{SD_{os}}{\sqrt{n}}$$

Međutim, kako je SD_{os} nepoznata, a uzorak zamenjuje osnovni skup, onda i standardnu devijaciju osnovnog skupa zamenjuje standardna devijacija uzorka.

Kako je varijansa osnovnog skupa veća od varijanse aritmetičkih sredina uzoraka, konačna formula za standardnu grešku glasi:

$SG = \frac{SDuz}{\sqrt{n-1}}$, gde je n=veličina uzorka, a sam izraz n-1 u statistici se naziva stepen slobode.

Standardna greška predstavlja varijabilnost aritmetičkih sredina uzoraka u odnosu na aritmetičku sredinu osnovnog skupa.

Standardna greška ima manju vrednost od stvarne standardne devijacije, jer su uzorci kompenzovali u određenoj meri varijabilnost.

Postavlja se pitanje kada treba koristiti standardnu devijaciju, a kada standardnu grešku?

Ova dva parametra su privedno slična, ali namena im je potpuno različita. Standardna devijacija se primenjuje kada je potrebno pokazati varijabilnost podataka, a standardna greška kada je potrebno ukazati na preciznost aritmetičke sredine uzorka u odnosu na aritmetičku sredinu osnovnog skupa.

Izraz n-1 naziva se u statistici stepen slobode. **Stepen slobode** predstavlja broj varijabli koje slobodne mogu da uzimaju vrednost, tj. da variraju. Stepen slobode se izračunava tako što se od ukupnog broja jedinica nekog uzorka, oduzme one jedinice koje ne mogu slobodno da uzimaju vrednost.

Interval poverenja

Formiranjem reprezentativnog uzorka stvara se mogućnost određivanja tj. izračunavanja parametara uzorka na osnovu kojih je moguće doneti ocenu o variranju posmatranog obeležja u osnovnom skupu. Intervali variranja nazivaju se intervali pouzdanosti ili poverenja.

Interval poverenja je brojčani interval u kome očekujemo da se nađu parametri osnovnog skupa na osnovu parametara uzorka za zadatu verovatnoću i njoj odgovarajuću z-vrednost.

Obrazac po kom se određuje interval poverenja je sledeći:

$$\bar{x}_{\mathrm{uz}} - z\,\mathrm{SG} < \bar{X}_{\mathrm{os}} < \bar{x}_{\mathrm{uz}} + z\,\mathrm{SG}$$

Interval poverenja je određen:

- verovatnoćom,
- z – vrednošću
- standardnom greškom.

Verovatnoća se označava kao **koeficijent pouzdanosti.** Najčešće se određuje za p=0,95 (95%) i p=0,99 (99%), ređe p=0,90 (90%). Za određenu verovatnoću moguće je odrediti z – vrednost koja predstavlja **faktor pouzdanosti.** Određivanje z-vrednosti na osnovu verovatnoće i obrnuto, sprovodi se uz pomoć tablice – Površina ispod normalne krive. Velikim slovima se u statistici označavaju parametri osnovnog skupa, a malim slovima parametri uzorka.

Ako u ovom slučaju pretpostavimo da je prosečna masa 16 novorođenčadi koju smo izračunali u prethodnom primeru, prost slučajan uzorak svih novorođenčadi u nekom porodilištu, postavlja se pitanje kolika je masa u tom porodilištu za verovatnoću od 95% i 99%?

$\bar{x}_{\mathrm{uz}} = 3{,}42$ SD=0,45 za verovatnoću 95% z =1,96	$\bar{x}_{\mathrm{uz}} = 3{,}42$ SD=0,45 za verovatnoću 99% z =2,58
$\mathrm{SG} = \dfrac{SD}{\sqrt{n-1}}$	$\mathrm{SG} = \dfrac{SD}{\sqrt{n-1}}$
$\mathrm{SG} = \dfrac{0{,}45}{\sqrt{16-1}} = 0{,}12$	$\mathrm{SG} = \dfrac{0{,}45}{\sqrt{16-1}} = 0{,}12$
$3{,}42 - 1{,}96 \cdot 0{,}12 < \bar{x}_{\mathrm{os}} < 3{,}42 + 1{,}96 \cdot 0{,}12$	$3{,}42 - 2{,}58 \cdot 0{,}12 < \bar{x}_{\mathrm{os}} < 3{,}42 + 2{,}58 \cdot 0{,}12$
$3{,}18 < X_{\mathrm{os}} < 3{,}66$	$3{,}11 < X_{\mathrm{os}} < 3{,}73$

Sa verovatnoćom od 95% tvrdimo da je prosečna masa novorođenčadi u intervalu od 3,18 – 3,66kg. Prosečna masa beba na rođenju je u intervalu 3,11 – 3,73kg za verovatnoću od 99%.

Na osnovu ovog primera lepo se može uočiti da preciznost i pouzdanost intervala stoje u obrnuto proporcionalnom odnosu. Tačnije što je veća preciznost, manja je pouzdanost, jer se dobija interval veće širine i obrnuto.

Zadaci za vežbanje

1. Ispitivana je telesna visina dečaka i devojčica u prvom razredu osnovne škole.

   Dečaci: n=150, $\bar{x}$=154cm, SD=11,3

   Devojčice: n=160, $\bar{x}$=142cm, SD=9,7

   Konstruisati interval poveranja za jednu i drugu populaciju za p=0,80, p=0,95, p=0,99.

2. Prosečno vreme oporavka nakon povrede ligamenata na Klinici za rehabilitaciju u Nišu je 33 dana i SD=4,5 dana.

Sa verovatnoćom od 95% i 99% odrediti koliko je prosečno lečenje svih pacijenata u Nišavskom okrugu?

3. U junskom ispitnom roku ispit iz statistike položilo je 150 studenata medicine i 50 studenata stomatologije. Prosečna ocena studenata medicine bila je 8,5 i SD=1,5, a studenata stomatologije 7,5 i SD=1,5.

Proceniti kolika je bila prosečna ocena svih studenta medicine i stomatologije u školskoj 2009/2010. godini?

4. Koliko se uzoraka veličine n=2, mogu dobiti uz vraćanje iz osnovnog skupa N=6 jedinica, a koliko uzoraka iste veličine bez vraćanja?

5. Iz uzorka od N=6 osoba kojima je izmeren sistolni arterijski pritisak, odaberi sve moguće uzorke veličine n=2 i konstruiši distribuciju njihove frekvencije, oko aritmetičke sredine osnovnog skupa. Sistolni pritisak je iznosio:

Osobe	Krvni pritisak
1	160
2	140
3	130
4	165
5	150
6	155

Na osnovu datih podataka dokaži:

prvo pravilo: $\bar{x}_{uz} = \bar{x}_{os}$

drugo pravilo: $SD_{os}^2 = n \cdot SD_{uz}^2$

da je $SD_{os} > SD_{uz}$

koliki se procenat uzoraka nalazi u intervalu: $\bar{x}_{os} \pm 1,65SG$

koliko je udaljen uzorak koji ima $\bar{x}$ =125 mm Hg od aritmetičke sredine osnovnog skupa.

6. Na osnovu podataka o apsentizmu (odsutnost sa posla zbog bolesti i povreda) u jednom preduzeću izvučen je uzorak od 150 radnika. Na osnovu uzorka izračunato je prosečno godišnje odsustvovanje sa posla po jednom radniku od 29 dana sa standardnom devijacijom od 6 dana. Oceniti sa verovatnoćom od P=0,9 prosečno vreme odsustvovanja za sve radnike.

7. Istraživač je želeo da ispita potrošnju lekova u jednom okrugu u odnosu na kategoriju osiguranja: radničko, zemljoradničko i neosigurana lica i to na uzorku od 1.400 lica. Navedeni okrug je imao 750000 stanovnika od kojih:

- radničko osiguranje	300.000
- zemljoradničko osiguranje	250.000
- neosigurana lica	200.000

Odredi proporcionalno učešće broja jedinica iz svakog stratuma u uzorku od 1.400 lica.

8. Ispitivana je visina holesterola u krvi kod seoske i gradske populacije i to na osnovu uzoraka:

a) gradsko stanovništvo: n=800; $\bar{x}$ =4,3 i SD=0,9

b) seosko stanovništvo: n=540; $\bar{x}$ =5,2 i SD=1,0

Konstruiši intervale pouzdanosti za jednu i drugu populaciju za P=0,999.