Inferencijalna statistika i intervali pouzdanosti

Od uzorka do zaključka o populaciji

Bojan B.

EFBL/FPN

20. april 2026.

Pregled predavanja

Od uzorka do populacije

Inferencijalna statistika — osnovna ideja

Inferencijalna statistika donosi **zaključke o populaciji** na osnovu podataka prikupljenih iz uzorka. Uvijek postoji neizvjesnost — ne posmatramo cijelu populaciju.

Tipična pitanja

- Koji procenat zaposlenih žena ima fakultetsku diplomu?
- Koliko je srednje vrijeme preživljavanja za pacijente s ovom vrstom raka?
- Koji udio birača podržava određenu politiku?
- Koja je prosječna plata u sektoru?

Parametar vs. statistika

Parametar — broj koji opisuje populaciju
Oznaka: μ (sredina), p (proporcija)

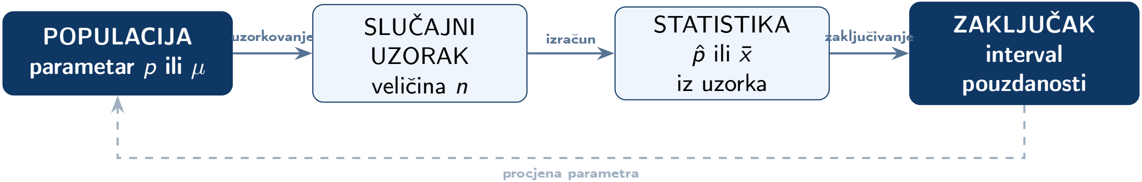
Statistika — broj izračunat iz uzorka
Oznaka: $\bar{x}$ (sredina), $\hat{p}$ (proporcija)

Statistiku *znamo* (iz podataka).
Parametar *procjenjujemo* (cilj analize).

Ključna razlika od deskriptivne statistike

Deskriptivna statistika *opisuje* uzorak koji imamo. Inferencijalna statistika *zaključuje* o populaciji koju nismo u potpunosti posmatrali.

Lanac zaključivanja: od podataka do zaključka



Šta znamo

- Statistiku iz uzorka ($\hat{p}$ ili $\bar{x}$)
- Veličinu uzorka n
- Da je uzorak slučajan (SRS)

Šta procjenjujemo

- Parametar populacije (p ili μ)
- S određenim **nivoom pouzdanosti** (95%, 99%...)
- U obliku **intervala**, ne tačke

Distribucija uzorkovanja proporcije

Formalna definicija

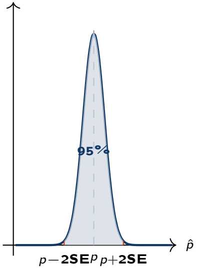
Distribucija uzorkovanja statistike je distribucija vrijednosti koje statistika uzima u *svim mogućim uzorcima* iste veličine iz iste populacije.

Distribucija uzorkovanja $\hat{p}$

Uzmimo SRS veličine n iz populacije s proporcijom p.

Kada je n dovoljno veliko:

- $\hat{p}$ prati **normalnu distribuciju**
- Sredina distribucije uzorkovanja: $\mu_{\hat{p}} = p$
- Standardna greška: $\text{SE} = \sqrt{\frac{p(1-p)}{n}}$



95% svih uzoraka daje $\hat{p}$ unutar $2 \cdot SE$ od p.

Primjer: Planovi apsolventa za postdiplomski studij

Postavljanje problema

Istina o populaciji: $p = 0{,}215$ (21,5% apsolventa planira postdiplomski).
Veličina uzorka: $n = 23.915$.

Izračun standardne greške

$$SE = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0{,}215 \times 0{,}785}{23.915}}$$

$$= \sqrt{\frac{0{,}16878}{23.915}} = \sqrt{0{,}00000706} = 0{,}00266$$

Pravilo 68-95-99,7:
95% uzoraka daje $\hat{p}$ između:
$0{,}215 - 2(0{,}00266) = 0{,}2097$
$0{,}215 + 2(0{,}00266) = 0{,}2203$

Šta ovo znači?

- SE je mali jer je n velik
- Skoro svi uzorci daju $\hat{p}$ blisko pravom p
- 95% uzoraka pada u raspon $[0{,}210;\ 0{,}220]$

Stabilnost SE

SE se malo mijenja s p:

p	SE
0,20	0,00259
0,215	0,00266

Šta je interval pouzdanosti?

Definicija

Interval pouzdanosti nivoa C za parametar je interval izračunat iz podataka uzorka koji “hvata” pravu vrijednost parametra u $C\%$ ponavljanih uzorkovanja.

Formula za proporciju

$$\hat{p} \pm z^* \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

gdje je:

- $\hat{p}$ — proporcija uzorka
- z^* — kritična vrijednost za nivo C
- n — veličina uzorka

Kritične vrijednosti z^*

Nivo pouzdanosti	z^*
90%	1,645
95%	1,960
99%	2,576

Koristimo $\hat{p}$ umjesto p jer p ne znamo. Za veliki n, ovo je dobra aproksimacija.

Napomena: *Koristimo $\hat{p}$ jer p ne znamo — to je upravo ono što procjenjujemo. Za veliki n ovo je dobra aproksimacija.*

Primjer: 95% interval pouzdanosti za proporciju

Podaci

NSSE anketa: $n = 23.915$ apsolventa, od kojih 5.038 planira postdiplomski.
$\hat{p} = 5038/23915 = 0{,}211$

Izračun 95% intervala

$$\mathrm{SE} = \sqrt{\frac{0{,}211 \times 0{,}789}{23915}} = \sqrt{0{,}000006970} = 0{,}00264$$

$$\hat{p} \pm z^* \cdot \mathrm{SE} = 0{,}211 \pm 1{,}960 \times 0{,}00264$$

$$= 0{,}211 \pm 0{,}0052$$

Interval: $[0{,}2058;\ 0{,}2162]$

Tumačenje

95% smo uvjereni da između **20,58%** i **21,62%** svih apsolventa u SAD i Kanadi planira postdiplomski studij.

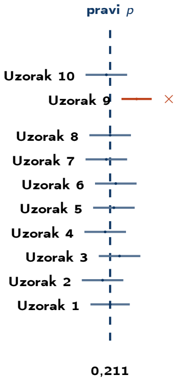
Drugim riječima: metoda koja daje ovakve intervale "hvata" pravi p u 95% slučajeva pri ponovljenom uzorkovanju.

Važno! Interval *ne* znači da postoji 95% vjerovatnoća da p leži u ovom konkretnom intervalu. Pravi p je fiksna vrijednost, nije slučajan.

Razumijevanje intervala pouzdanosti

Ključna ideja: šta bi se desilo pri ponovljenom uzorkovanju?

Svaki uzorak daje *drugačiji* interval. Nivo pouzdanosti 95% znači da bi 95 od 100 takvih intervala "uhvatilo" pravi parametar.



Dva konkretna uzorka

Uzorak 1: $\hat{p} = 0{,}211$

$\Rightarrow$ interval: $0{,}211 \pm 0{,}0053 = [0{,}206;\ 0{,}216]$

Uzorak 2: $\hat{p} = 0{,}208$

$\Rightarrow$ interval: $0{,}208 \pm 0{,}0052 = [0{,}203;\ 0{,}213]$

Oba intervala su različita, ali oba "hvataju" pravi $p \approx 0{,}215$.

Pouzdanost metode, ne intervala Govorimo o pouzdanosti **metode**, ne o pouzdanosti jednog konkretnog intervala. 95% intervala napravljenih ovom metodom sadrži pravi parametar.

Različiti nivoi pouzdanosti

Opšta formula s kritičnom vrijednošću z^*

Za bilo koji nivo pouzdanosti C između 0 i 1, postoji z^* takav da normalna distribucija ima vjerovatnoću C unutar z^* standardnih devijacija od sredine.

$$\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

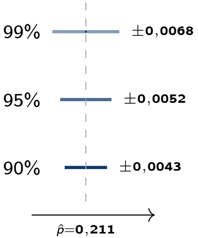
Primjer: 99% interval (NSSE)

$n = 23.915, \quad \hat{p} = 0{,}211, \quad z^* = 2{,}576$

$\hat{p} \pm 2{,}576 \times 0{,}00264$
$= 0{,}211 \pm 0{,}0068$

Interval: $[0{,}2042;\ 0{,}2178]$

99% smo uvjereni da između 20,42% i 21,78% apsolventa planira postdiplomski.



Viša pouzd. $\Rightarrow$ širi int. $\Rightarrow$ manje prec.

Distribucija uzorkovanja srednje vrijednosti $\bar{x}$

Teorija

Odaberemo SRS veličine n iz populacije sa sredinom μ i SD σ. Neka je $\bar{x}$ srednja vrijednost uzorka.

- Distribucija uzorkovanja $\bar{x}$ je **aproximativno normalna** za velik n
- Sredina distribucije uzorkovanja: $\mu_{\bar{x}} = \mu$ i standardna greška: $\text{SE}_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}}$

Dvije važne posljedice

1. Manja varijabilnost:
Sredina više opservacija je *manje varijabilna* od pojedinačnih opservacija.

2. Normalnost:
Distribucija sredine je normalija od distribucije pojedinačnih opservacija.

Standardna greška se smanjuje s n

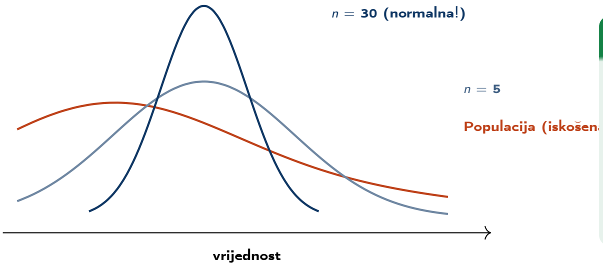
n	$\text{SE} = \sigma/\sqrt{n}$	vs. $n = 1$
1	σ	$\times 1$
4	$\sigma/2$	$\times 0{,}50$
25	$\sigma/5$	$\times 0{,}20$
100	$\sigma/10$	$\times 0{,}10$

Da bismo prepolovili SE, trebamo 4× veći uzorak.

Centralna granična teorema

Centralna granična teorema (CGT)

Kako uzimamo sve više i više nasumičnih opservacija iz *bilo koje* populacije, distribucija srednje vrijednosti tih opservacija se **sve više približava normalnoj distribuciji**.



Praktično pravilo

- $n \geq 15$: adekvatno za većinu distribucija (bez ekstremnih outliera)
- $n \geq 40$: dovoljna za jasno iskošene distribucije bez outliera
- Veći n = sigurnija normalna aproksimacija

Zašto je CGT važna?

Čak i kada ne znamo oblik distribucije populacije, za dovoljno velik n možemo koristiti normalne metode za intervale pouzdanosti!

Interval pouzdanosti za srednju vrijednost μ

Formula

Odaberemo SRS veličine n iz populacije sa nepoznatom sredinom μ.
Neka je $\bar{x}$ sredina uzorka, s standardna devijacija uzorka.
Aproximativni interval pouzdanosti nivoa C za μ:

$$\bar{x} \pm z^* \cdot \frac{s}{\sqrt{n}}$$

Analogija s proporcijom

	Proporcija	**Sredina**
Parametar	p	μ
Statistika	$\hat{p}$	$\bar{x}$
SE	$\sqrt{\hat{p}(1-\hat{p})/n}$	$s/\sqrt{n}$
Interval	$\hat{p} \pm z^* \cdot \text{SE}$	$\bar{x} \pm z^* \cdot \text{SE}$

Napomene

- Ne znamo σ (populacijska SD), pa koristimo s (SD uzorka) - valjano za velik n
- Za mali n koristimo t-distribuciju (napredni kurs)
- Recept vrijedi samo za SRS!

Primjer: NAEP test iz matematike

Podaci

NAEP 2015: $n = 13.200$ maturanta, $\bar{x} = 152$ (rezultat), $s = 34$, raspon 0–300.

Izračun 95% intervala

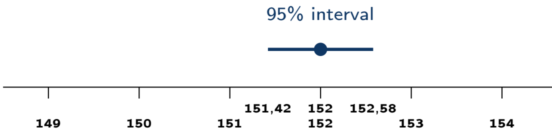
$$SE = \frac{s}{\sqrt{n}} = \frac{34}{\sqrt{13200}} = \frac{34}{114{,}9} = 0{,}296$$

$$\bar{x} \pm z^* \cdot SE = 152 \pm 1{,}960 \times 0{,}296$$

$$= 152 \pm 0{,}58$$

Interval: $[151{,}42;\ 152{,}58]$

95% smo sigurni da je srednji rezultat svih maturanta između **151,42** i **152,58**.



Uočite preciznost

Margina greške je samo $\pm 0{,}58$ boda!
To je jer je $n = 13.200$ veoma velik.
Veći $n \Rightarrow$ uži interval $\Rightarrow$ preciznija procjena.

Sažetak: Sve u jednoj tablici

Situacija	**Parametar**	**Interval**	**SE**
Proporcija	p (nepoznat)	$\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$	$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
Sredina	μ (nepoznat)	$\bar{x} \pm z^* \frac{s}{\sqrt{n}}$	$\frac{s}{\sqrt{n}}$

Uvjeti za valjanost

- Uzorak je **SRS** (slučajni)
- Veličina uzorka **dovoljno velika** ($n \geq 15$ za sredinu, $n\hat{p} \geq 10$ za proporciju)
- Populacija znatno veća od uzorka

Česte greške u tumačenju

- "95% šanse da pravi parametar leži u intervalu" — pogrešno (parametar nije slučajan)
- "95% podataka je u intervalu" — pogrešno (interval nije za podatke)
- Ispravno: 95% je *pouzdanost metode*

Praktično pitanje 1: Podrška politici

Scenarij

Anketiramo SRS od $n = 1.500$ odraslih građana. Od njih, $n_{\text{da}} = 825$ podržava određenu ekonomsku politiku.
$\hat{p} = 825/1500 = 0{,}550$

Zadaci

1. Izračunaj 95% interval pouzdanosti za p.
2. Izračunaj 99% interval pouzdanosti za p.
3. Može li se tvrditi da *više od polovine* građana podržava politiku? Zašto?
4. Kako bi se rezultat promijenio da je $n = 400$?

Rješenja

$\text{SE} = \sqrt{\frac{0{,}55 \times 0{,}45}{1500}} = \sqrt{0{,}0001650} = 0{,}01285$

(1) $0{,}550 \pm 1{,}960 \times 0{,}01285 = 0{,}550 \pm 0{,}025$
Interval: $[0{,}525;\ 0{,}575]$

(2) $0{,}550 \pm 2{,}576 \times 0{,}01285 = 0{,}550 \pm 0{,}033$
Interval: $[0{,}517;\ 0{,}583]$

(3) Da — oba intervala su *iznad* 0,50.

(4) $\text{SE} = \sqrt{0{,}55 \times 0{,}45/400} = 0{,}0249$
Mnogo širi interval: $0{,}550 \pm 0{,}049$

Praktično pitanje 2: Prosječna plata

Scenarij

SRS od $n = 400$ zaposlenih u sektoru IT. Prosječna plaća uzorka: $\bar{x} = 2.850\,\text{KM}$, standardna devijacija uzorka: $s = 520\,\text{KM}$.

Zadaci

1. Izračunaj standardnu grešku $\bar{x}$.
2. Izračunaj 95% interval pouzdanosti za μ.
3. Tumači interval u kontekstu.
4. Medijski naslov kaže "prosječna IT plaća je 2.900 KM". Je li ovo kompatibilno s vašim intervalom?

Rješenja

(1) $\text{SE} = \frac{520}{\sqrt{400}} = \frac{520}{20} = 26\,\text{KM}$

(2) $2850 \pm 1{,}960 \times 26 = 2850 \pm 50{,}96$
Interval: $[2799;\ 2901]\,\text{KM}$

(3) 95% smo sigurni da je prosječna IT plaća između 2.799 i 2.901 KM.

(4) 2.900 KM *leži unutar* intervala $[2799; 2901]$ — tvrdnja iz naslova je kompatibilna s našim podacima.

Praktično pitanje 3: Veličina uzorka i pouzdanost

Zadatak za razumijevanje

Istraživač želi 95% interval pouzdanosti za proporciju p koji ima marginu greške **ne veću od 0,03** (3 procentna poena). Koliki uzorak je potreban?

Rješenje

Margina greške: $z^* \cdot \text{SE} \leq 0{,}03$

$$1{,}960\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq 0{,}03$$

Najgori slučaj: $\hat{p} = 0{,}5$ (maksimalni SE)

$$1{,}960 \times \sqrt{\frac{0{,}25}{n}} \leq 0{,}03$$

$$\sqrt{\frac{0{,}25}{n}} \leq \frac{0{,}03}{1{,}960} = 0{,}01531$$

$$n \geq \frac{0{,}25}{0{,}01531^2} = \frac{0{,}25}{0{,}000234} \approx 1068$$

Opća formula za veličinu uzorka

$$n \geq \left(\frac{z^*}{ME}\right)^2 \hat{p}(1-\hat{p})$$

Za nepoznato $\hat{p}$, koristimo $\hat{p} = 0{,}5$ (daje najkonzervativniju — najveću — procjenu n).

Zaključak

Trebamo barem $n = 1068$ ispitanika za marginu greške $\leq 3\%$ pri 95% pouzdanosti.

Hvala na pažnji!

Pitanja i diskusija

"Nikada ne možemo biti 100% sigurni — ali možemo biti 95%."

— *osnovna ideja intervala pouzdanosti*