1. Populacija vs uzorak

Populacija je sve. Uzorak je deo. Statistika pokušava da iz dela zaključi o svemu.

Primer 1:

Ako želite da znate **prosečnu visinu studenata na fakultetu**:

- **Populacija** → svi studenti na fakultetu
- **Uzorak** → 100 studenata koje ste izabrali i izmerili

Ne merite sve studente jer može da traje predugo, skupo je I često je nepotrebno. Zato uzmete **uzorak** i **procijenite** populaciju.

Primer 2 (lako se pamti 😊):

- **Populacija** = ceo lonac supe
- **Uzorak** = jedna kašika supe koju probate

Ako je kašika dobro izmešana → dobijate dobru procenu celog lonca.

Ako nije izmešana → dobijate pogrešan zaključak.

Ovde se prirodno uvodi pojam **reprezentativnost uzorka**. Setite se primera sa predavanja u vezi istraživanja mnjenja uoči političkih izbora**.**

2. Unbiased (nepristrasno)

* Deo gde pričamo o proseku na strani 5

Kaže se **unbiased (nepristrasna procena)** zato što **u proseku ne grešimo sistematski**.

Formalno, procena je **unbiased** ako važi:

$$E[\hat{\mu}] = \mu$$

To znači:

- Ako bismo **mnogo puta uzimali uzorak**
- Svaki put izračunali **srednju vrednost uzorka** ($\hat{\mu}$)
- I onda uzeli **prosek svih tih procena**

→ dobili bismo **tačno pravu vrednost populacije** (μ)

Primer sa predavanja

Zamislite da:

- Pravi prosek visine studenata = **175 cm**
- Uzmemo jedan uzorak → dobijemo **174 cm**, drugi uzorak → **176 cm**, treći → **175.5 cm,** četvrti → **174.8 cm**

Nekad pogodimo više, nekad manje — ali **nema sistematskog pomeranja**. Drugim rečima, grešimo ponekad, ali **ne grešimo uvek u istom smeru.** To je **unbiased**.

biased procena

Ako bismo uvek birali:

- samo odbojkaše
- samo studente prve godine
- samo muškarce

→ prosek bi **stalno bio veći ili manji od stvarnog.** To je **biased procena** (pristrasna).

3. t-vrednost

U linearnoj regresiji:

- Želimo da proverimo da li postoji veza između **X i Y**
- To znači da testiramo da li je **nagib (β_1)** različit od nule

Hipoteze:

- **H_o:** $\beta_1 = 0$ → nema veze između X i Y
- **H_1:** $\beta_1 \neq 0$ → postoji veza između X i Y

t-statistiku računamo zato što želimo da procenimo **da li je uočena veza stvarna ili samo posledica slučajnosti**.

t-statistika nam zapravo govori: **Koliko je procena udaljena od nule u odnosu na njenu nesigurnost.**

Veliko t => verovatno postoji veza, a ako je malo t => verovatno nema veze

Na primer, ako je:

- $\beta_1 = 2.5$
- Standardna greška = 0.5

t = 2.5 / 0.5 = 5 → **veliko → verovatno postoji veza**

Ali ako je:

- $\beta_1 = 2.5$
- Standardna greška = 5

t = 2.5 / 5 = 0.5 → **malo → verovatno nema veze**

t-statistika meri signal u odnosu na šum:

β_1 = signal, SE = šum

t = signal / šum

Ako je signal mnogo veći od šuma → postoji veza, tj. meri se **koliko je procena daleko od nule** u odnosu na šum

Dodatno, Ako **zaista ne postoji veza** ($\beta_1 = 0$), koliko je **verovatno** da dobijemo ovako veliki t **samo slučajno**? Zato nam treba **p-vrednost**.

p-vrednost = verovatnoća da dobijemo ovako veliki (ili veći) t ako veze zapravo nema.

Primer:

Neka je t = 2.8 i p-value = 0.006

To znači:

Ako nema veze između X i Y, samo u 0.6% slučajeva bismo dobili ovako veliki t. Zato kažemo:
→ Malo verovatno da je slučajno
→ Verovatno postoji veza

Napomena: t-statistika je broj koji izračunamo, a t-raspodela nam govori koliko je taj broj verovatan. Drugim rečima:

- t-statistika → ono što dobijemo iz podataka
- t-raspodela → referenca da vidimo da li je to "veliko" ili "malo"