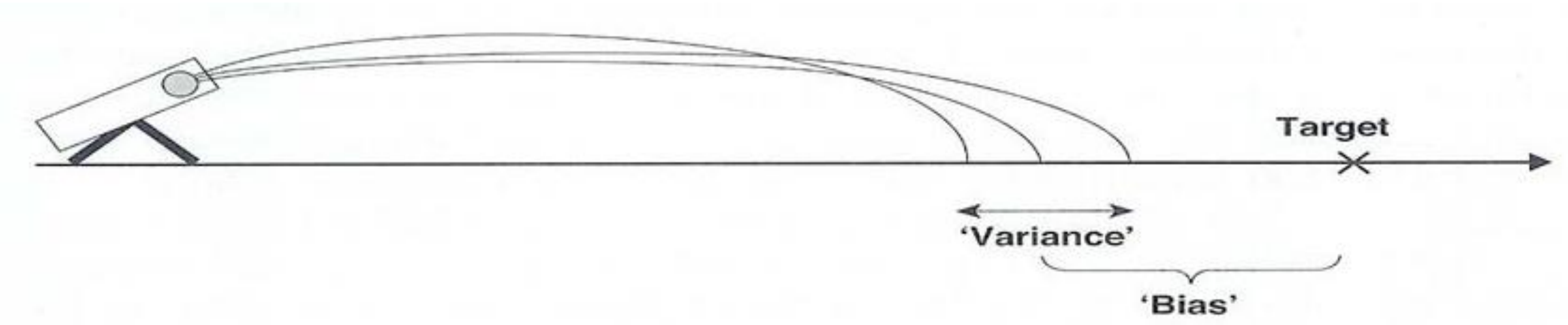
Estimacija parametara modela

- Neka je $\hat{\theta}$ ocena parametara θ
- Ocena $\hat{\theta}$ je *nepristrasna* (*centrirana*) ako važi

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

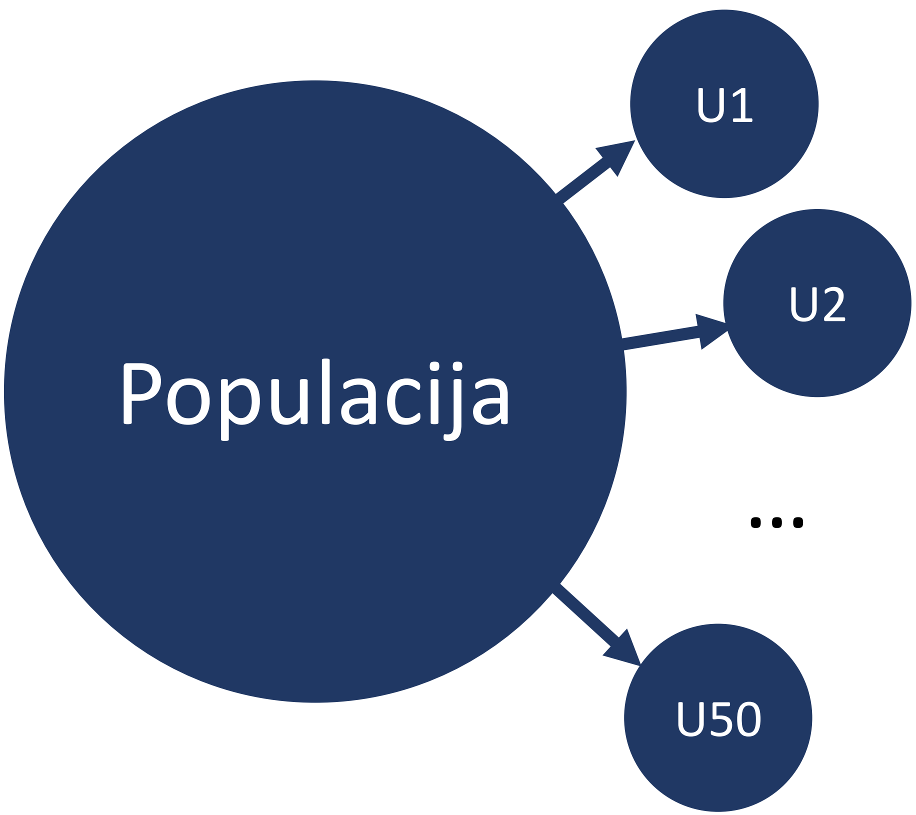
- Razlika $E[\hat{\theta}] - \theta$ je *sistematsko odstupanje* (*bias*) ocene $\hat{\theta}$
- Nepristrasna ocena koja ima manju varijansu je *bolja* od nepristrasne ocene koja ima veću varijansu



Pristrastnost – primer

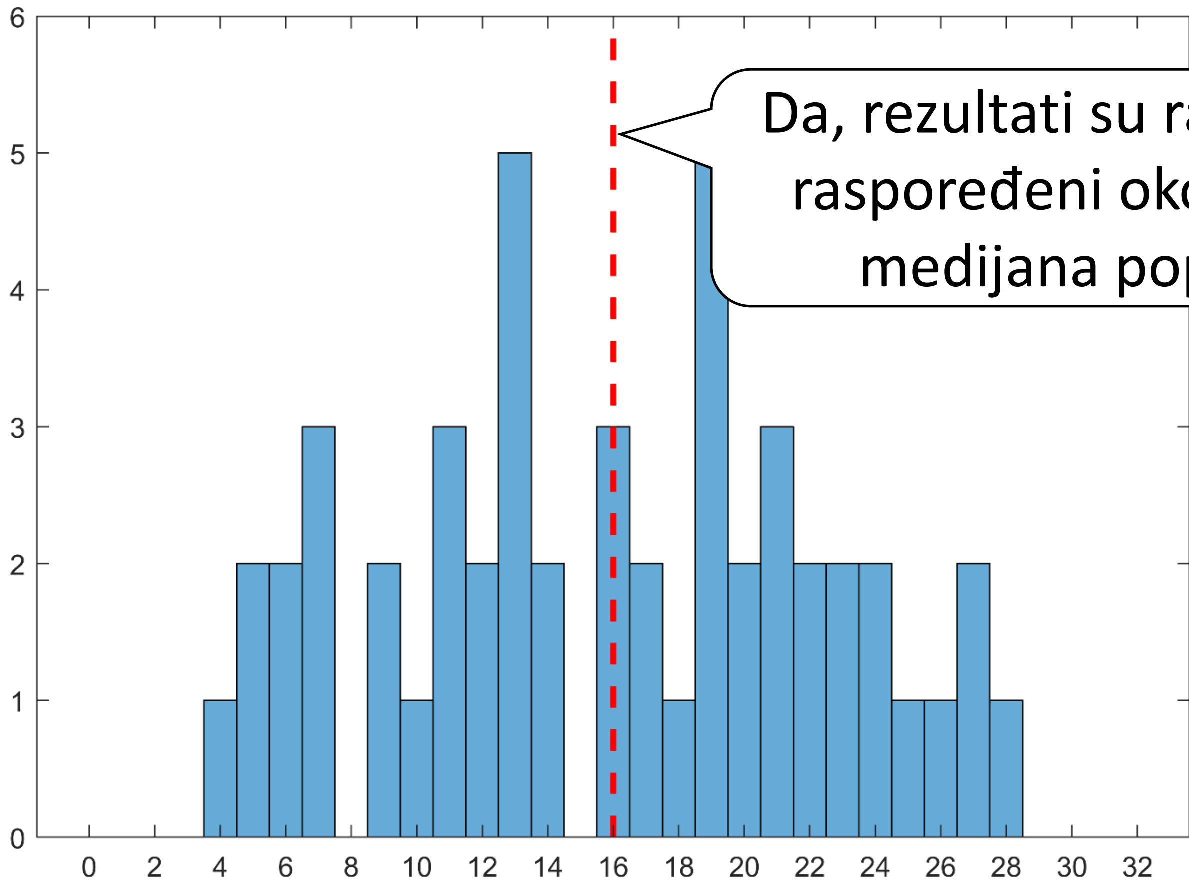
Da li je medijana uzorka nepristrastna ocena medijane populacije?

- Loptice označene brojevima 0-32
- Izmešamo u bubnju, izvučemo 5 loptica i izračunamo medijanu uzorka
- Ponovimo eksperiment 50 puta



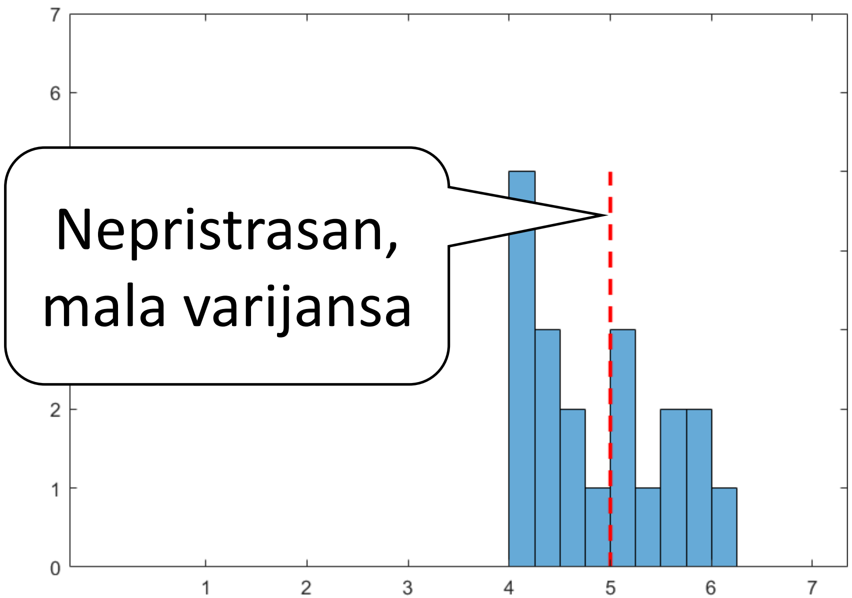
Pristrastnost – primer

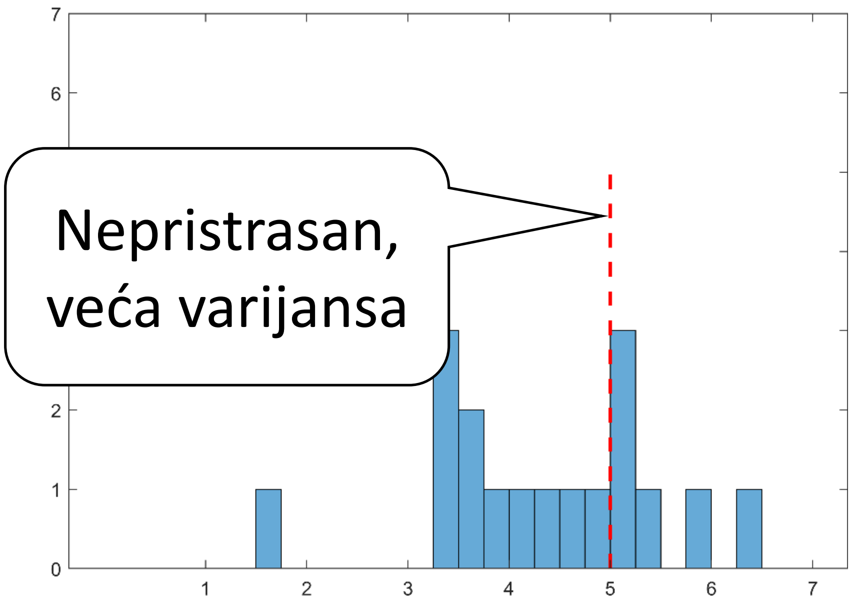
Da li je medijana uzorka nepristrastna ocena medijane populacije?

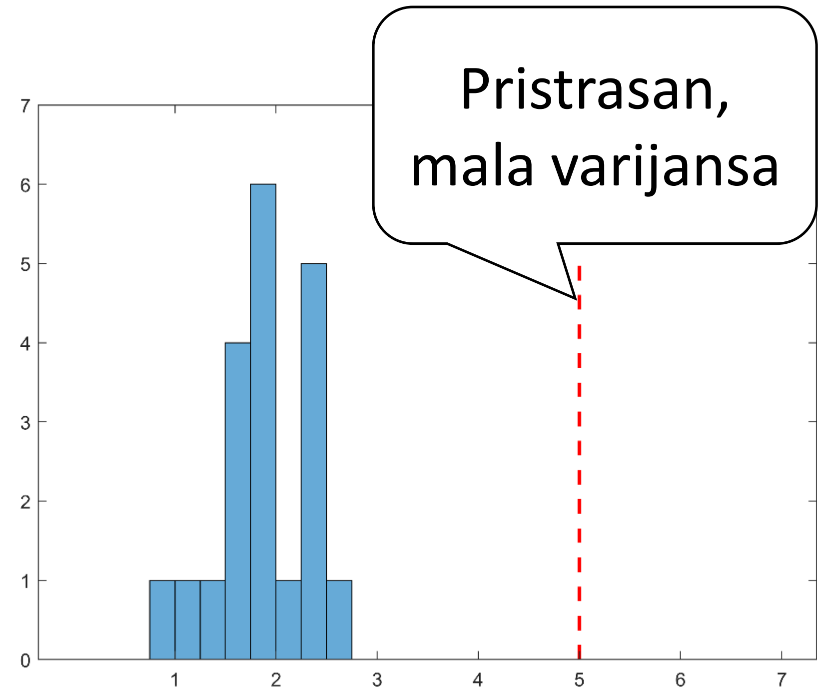


Da, rezultati su ravnomerno raspoređeni oko stvarnog medijana populacije

Pristrastnost – primer







Estimacija parametara normalne raspodele

- μ_{ML}, σ_{ML} su *ocene maksimalne verodostojnosti* za stvarne vrednosti μ i σ

- U slučaju da su podaci IID iz normalne raspodele:
  - μ_{ML} odgovara srednjoj vrednosti uzorka (nepristransa ocena)
  - σ_{ML} ne odgovara varijansi uzorka (nije nepristrasna ocena)

- ML pristup sistematski podcenjuje varijansu distribucije. Ovo je primer fenomena koji se zove *sistematsko odstupanje* (*bias*)

- Povezan je sa problemom *overfitting*-a koji smo videli kod fitovanja polinomijalne krive

Estimacija parametara normalne raspodele

- Ocena $\hat{\theta}$ je *nepristrasna* (*centrirana*) ako važi $E[\hat{\theta}] = \theta$

$$E[\mu_{ML}] = E\left[\frac{1}{N}\sum_{n=1}^{N} x^{(n)}\right] = \frac{1}{N}\sum_{n=1}^{N} E[x^{(n)}] = \frac{1}{N}\sum_{n=1}^{N} \mu = \mu$$

$E[\alpha x + \beta y] = \alpha E[x] + \beta E[y]$

μ kod normalne raspodele

- Pošto je $E[\mu_{ML}] = \mu$, μ_{ML} je nepristrasna ocena (*unbiased estimate*) parametra μ
- U proseku, ML ocena će rezultovati korektnom srednjom vrednošću

Estimacija parametara normalne raspodele

$$E[\sigma_{ML}^2] = E\left[\frac{1}{N}\sum_{n=1}^{N}(x_n - \mu_{ML})^2\right] =$$

$E[\alpha x + \beta y] = \alpha E[x] + \beta E[y]$

$$= E\left[\frac{1}{N}\sum_{n=1}^{N}(x_n)^2 - 2\mu_{ML}\underbrace{\frac{1}{N}\sum_{n=1}^{N}x_n}_{\mu_{ML}} + \frac{1}{N}N\mu_{ML}^2\right] =$$

$$= E\left[\frac{1}{N}\sum_{n=1}^{N}(x_n)^2 - \mu_{ML}^2\right] = \frac{1}{N}\sum_{n=1}^{N}E[x_n^2] - E[\mu_{ML}^2] =$$

$$= \frac{1}{N}\sum_{n=1}^{N}(\sigma^2 + \mu^2) - \left(\frac{\sigma^2}{N} + \mu^2\right) = \frac{(N-1)\sigma^2}{N}$$

var[x]=σ^2=E[x^2]-μ^2 $\quad$ $var[\mu] = \sigma^2/N = E[\mu^2] - \mu^2$

https://onlinecourses.science.psu.edu/stat414/node/167

Estimacija parametara normalne raspodele

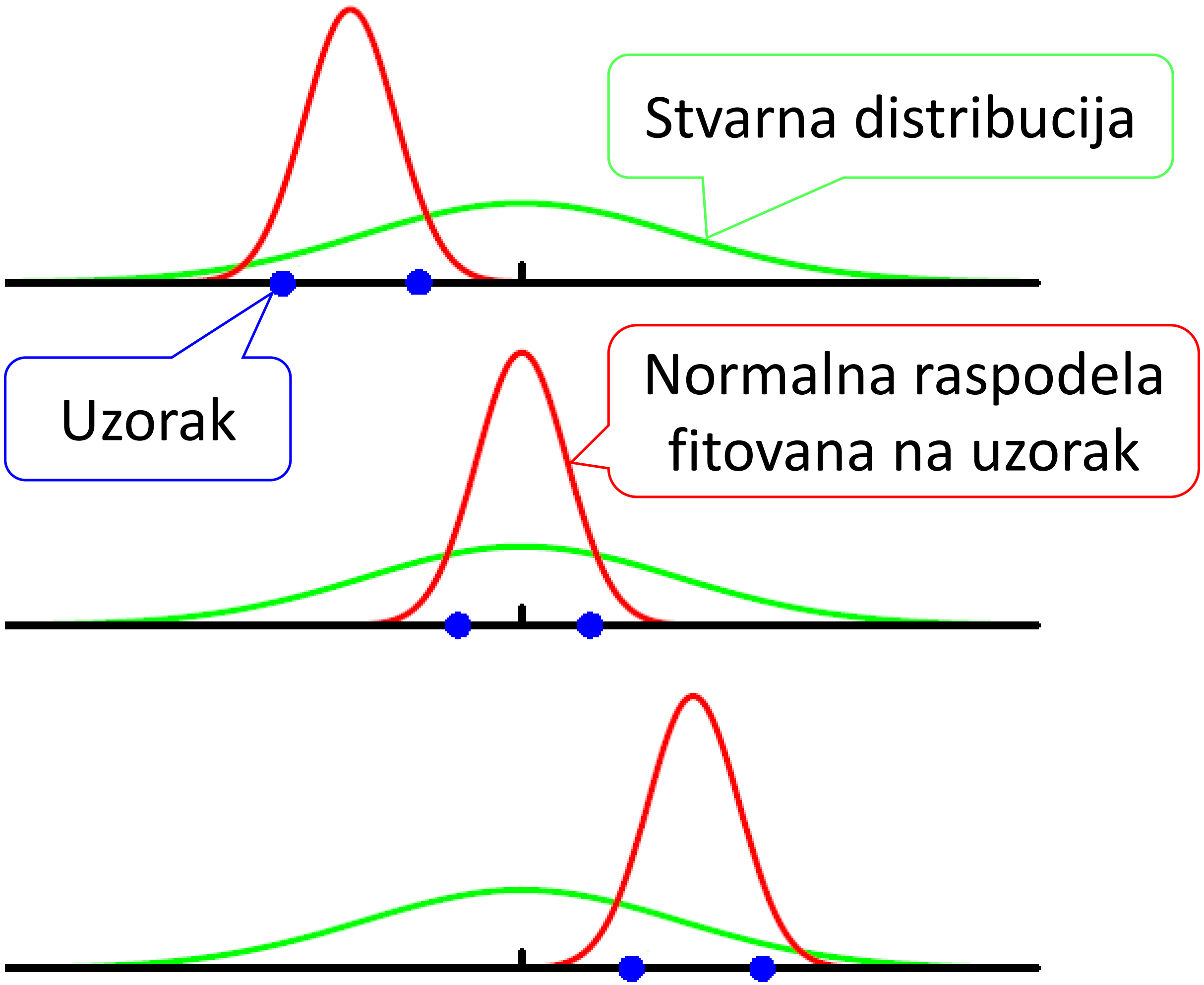
- Pošto je $E[\sigma_{ML}^2] \neq \sigma^2$, σ_{ML}^2 je pristrasna ocena (*biased estimate*) parametra σ^2

- U proseku, ova ocena će podceniti pravu varijansu za faktor $(N-1)/N$

- Nepristrasna ocena σ^2:

$$\tilde{\sigma}^2 = \frac{N}{N-1}\sigma_{ML}^2 = \frac{1}{N-1}\sum_{n=1}^{N}(x_n - \mu_{ML})^2$$

Estimacija parametara normalne raspodele



Ako bi se μ_{ML} uprosečio na ova tri skupa podataka, dobili bismo μ

σ^2 je sistematski podcenjena jer se računa relativno u odnosu na prosek uzorka (μ_{ML}), a ne u odnosu na μ