Utjecaj distribucije podataka na vjerojatnost pokrivanja t-intervala pouzdanosti za sredinu populacije

Projekt iz kolegija Računarska statistika (zadatak 5.)

Luka Šimek

Prirodoslovno-matematički fakultet — Matematički odsjek
Sveučilište u Zagrebu

11. siječnja 2025.

Sadržaj

Pojmovi

$$\gamma_1 = \mathbb{E}\left(\frac{X - \mathbb{E}X}{\mathrm{Var}X}\right)^3 \quad \text{i} \quad \gamma_2 = \mathbb{E}\left(\frac{X - \mathbb{E}X}{\mathrm{Var}X}\right)^4 - 3$$

Ako je $X \sim N(\mu, \sigma)$ i $X_1, X_2, \ldots, X_n$ slučajni uzorak, tada je

$$\frac{\overline{X}_n - \mu}{S_n}\sqrt{n} \sim t(n-1),$$

pri čemu je S_n uzoračka varijanca. Iz toga dobivamo pouzdani interval (t-interval) pouzdanosti $1 - \alpha$ za sredinu populacije μ kao

$$\overline{X}_n \pm \frac{S_n}{\sqrt{n}} t_{\frac{\alpha}{2}, n-1}.$$

Pitanje: vrijedi li slični rezultat ako X nije normalno distribuirana?

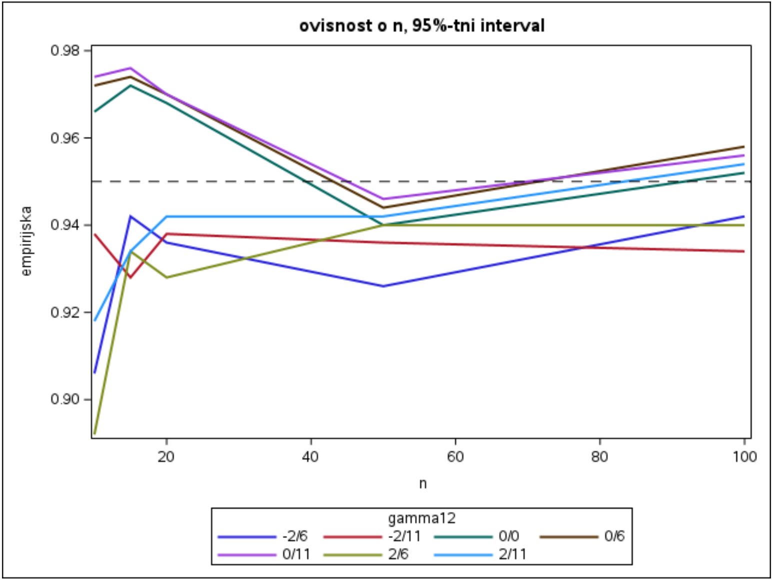
Parametri

MC studiju provodimo po kombinacijama faktora:

- duljina uzorka $n = 10, 15, 20, 50, 100$
- $\gamma_1 = -2, 0, 2$
- $\gamma_2 = 0, 6, 11$
- $\mu = 0, \sigma = 5$

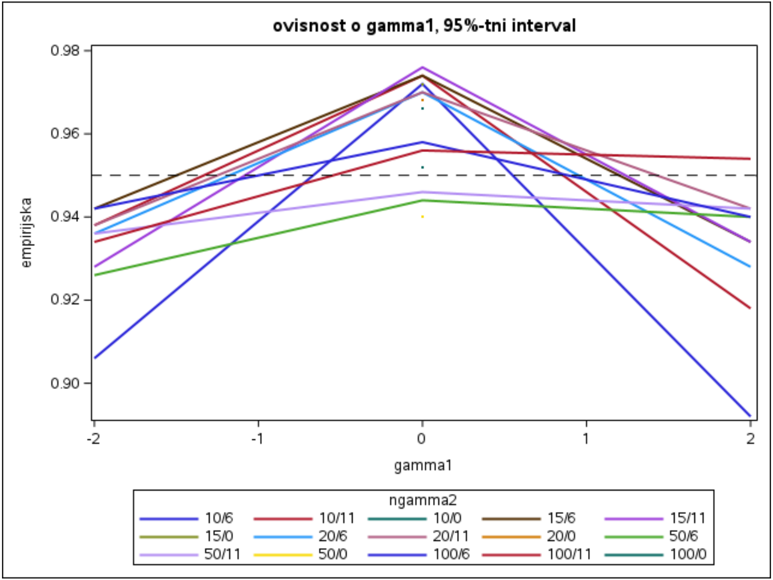
Za svaku trojku (n, γ_1, γ_2) simuliramo 500 uzoraka duljine n. Na temelju njih procjenjujemo stvarnu vjerojatnost pokrivanja nominalno 90%-p.i. i 95%-p.i.

Ovisnost o *n*

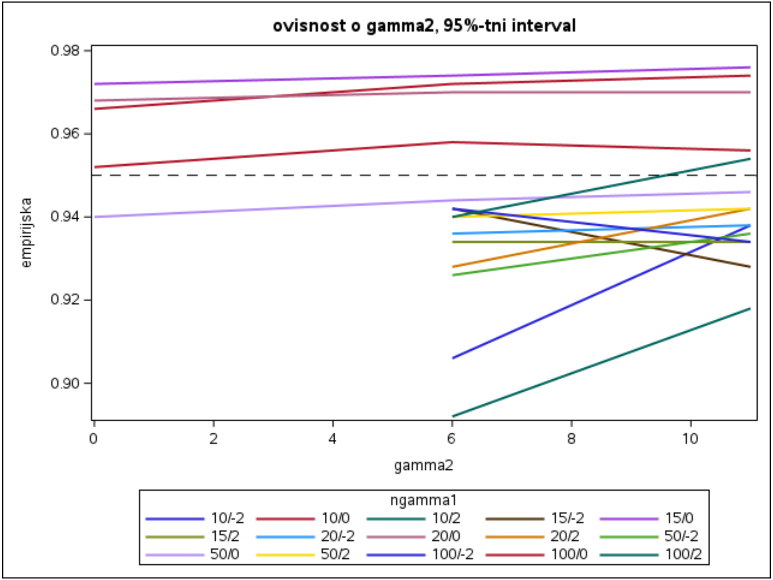


ovisnost o n, 95%-tni interval
0.98
0.96
0.94
0.92
0.90
empirijska
20
40
60
80
100
n
gamma12
-2/6
-2/11
0/0
0/6
0/11
2/6
2/11

Ovisnost o γ_1



Ovisnost o γ_2



Zaključci

- problematične su nesimetrične distribucije, naročito s malim uzorcima
- za simetrične spljoštene distribucije rezultati su slični kao za normalnu