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

Projekt iz kolegija Računarska statistika (zadatak 5.)

Luka Šimek

Zagreb, 11. siječnja 2025.

1 Pojmovi

Koeficijente asimetrije γ_1 i spljoštenosti γ_2 definiramo kao redom treći i četvrti moment slučajne varijable odnosno uzorka:

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

pri čemu se u definiciji od γ_2 dodatno oduzima 3 jer je to (ukupna) spljoštenost normalne distribucije. Na taj način distribucije s negativnim γ_2 imaju manju, a s pozitivnim veću spljoštenost u odnosu na normalnu distribuciju. Iz gornjih formula se vidi da je γ_1 po apsolutnoj vrijednosti veći kad je distribucija više „nagnuta", dok je γ_2 veći kad distribucija ima „deblje repove". U slučaju normalne distribucije obje veličine iznose nula.

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

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

2 Zadatak

U ovom projektu želimo ispitati utjecaj triju faktora: duljine uzorka n, koeficijenta asimetričnosti (*skewness*) γ_1 i koeficijenta spljoštenosti (*excess kurtosis*) γ_2 na stvarnu vjerojanost da (nominalno) 90% odn. 95% pouzdani interval sadrži sredinu populacije. U slučaju da razdioba podataka nije normalna ($\gamma_1 = 0$ i $\gamma_2 = 0$) stvarna vjerojanost ne mora odgovarati nominalnoj.

Konkretno, ispitujemo kombinacije sa sljedećim mogućnostima:

- $n = 10, 15, 20, 50, 100,$
- $\gamma_1 = -2, 0, 2$ i
- $\gamma_2 = 0, 6, 11$[1]

[1] u tekstu zadatka navodi se i $\gamma_2 = -3$ no to je generalno nemoguće pa nije navedeno

što nakon uzimanja u obzir danih nejednakosti daje sedam mogućih parova (γ_1, γ_2) i tridesetpet mogućih trojki (n, γ_1, γ_2). Za svaku takvu trojku generiramo 500 uzoraka duljine n, sredine $\mu = 0$, standardne devijacije $\sigma = 5$ i asimetričnosti i spljoštenosti γ_1 i γ_2. U svakoj od 500 replikacija ispitujemo pripadnost μ izračunatim pouzdanim intervalima i pomoću svih 500 rezultata dolazimo do empirijski dobivene vjerojatnosti pokrivanja. Za svaku trojku koristimo početni seed od $112025 + n\gamma_1\gamma_2$.

Kopija teksta zadatka priložena je na kraju ovog dokumenta.

3 Rezultati

Cjelokupne rezultate simulacije može se vidjeti u tablicama 1 i 2. Rezultate vizualiziramo na grafovima 3, 4, 5, 6, 7, 8 na kojima prikazujemo ovisnost vjerojanosti pokrivanja o jednom od triju faktora uz fiksiranje preostala dva.

Primjećujemo sljedeće:

- Najveća odstupanja vidimo za najmanju veličinu uzorka $n = 10$ i nesimetrične distribucije ($\gamma_1 = \pm 2$) — tada se svakako interval nebi trebao korisiti. Odstupanje se smanjuje s povećanjem n, no najdrastičnije za $n = 15$ i manje kasnije.
- U velikoj većini slučajeva vidimo da vjerojatnost pokrivanja pada kad distribucija nije simetrična zbog čega grafovi imaju karakteristični oblik „krovića".
- Ne uočavamo jasnu vezu između vjerojatnosti pokrivanja i γ_2; razlike su male kod simetričnih distribucija, a kod asimetričnih rezultati variraju.

Da rezimiramo, možemo reći da su najproblematičnije asimetrične distribucije, a pogotovo kad je uzorak mali. S druge strane, spljoštenost se nije pokazala problemom — na grafovima 3 i 4 vidimo da su rezultati za simetrične distribucije, čak i u slučaju najveće spljoštenosti, vrlo slični onima normalne distribucije.

4 Tablice i grafovi

1

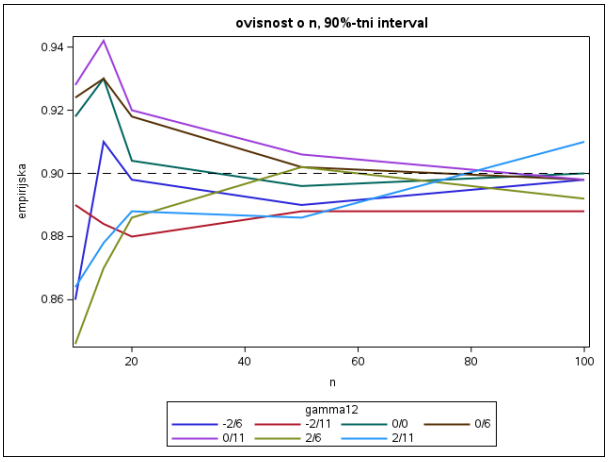
Obs	n	gamma1	gamma2	empirijska	nominalna
1	10	-2	6	0.860000	0.9
2	10	-2	11	0.890000	0.9
3	10	0	0	0.918000	0.9
4	10	0	6	0.924000	0.9
5	10	0	11	0.928000	0.9
6	10	2	6	0.846000	0.9
7	10	2	11	0.864000	0.9
8	15	-2	6	0.910000	0.9
9	15	-2	11	0.884000	0.9
10	15	0	0	0.930000	0.9
11	15	0	6	0.930000	0.9
12	15	0	11	0.942000	0.9
13	15	2	6	0.870000	0.9
14	15	2	11	0.878000	0.9
15	20	-2	6	0.898000	0.9
16	20	-2	11	0.880000	0.9
17	20	0	0	0.904000	0.9
18	20	0	6	0.918000	0.9
19	20	0	11	0.920000	0.9
20	20	2	6	0.886000	0.9
21	20	2	11	0.888000	0.9
22	50	-2	6	0.890000	0.9
23	50	-2	11	0.888000	0.9
24	50	0	0	0.896000	0.9
25	50	0	6	0.902000	0.9
26	50	0	11	0.906000	0.9
27	50	2	6	0.902000	0.9
28	50	2	11	0.886000	0.9
29	100	-2	6	0.898000	0.9
30	100	-2	11	0.888000	0.9
31	100	0	0	0.900000	0.9
32	100	0	6	0.898000	0.9
33	100	0	11	0.898000	0.9
34	100	2	6	0.892000	0.9
35	100	2	11	0.910000	0.9

Slika 1: Tablica s rezultatima, nominalno 90%-p.i.

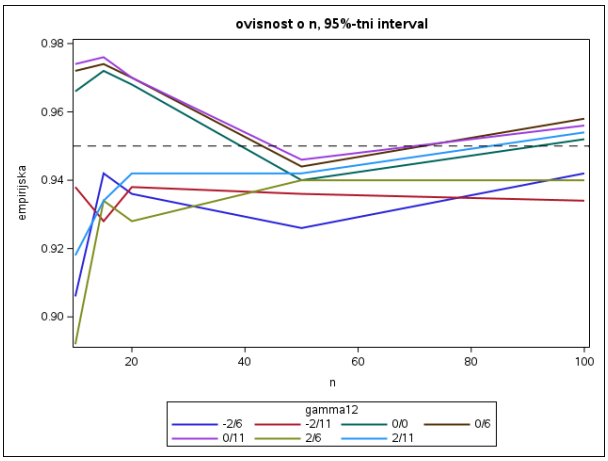
1

Obs	n	gamma1	gamma2	empirijska	nominalna
36	10	-2	6	0.906000	0.95
37	10	-2	11	0.938000	0.95
38	10	0	0	0.966000	0.95
39	10	0	6	0.972000	0.95
40	10	0	11	0.974000	0.95
41	10	2	6	0.892000	0.95
42	10	2	11	0.918000	0.95
43	15	-2	6	0.942000	0.95
44	15	-2	11	0.928000	0.95
45	15	0	0	0.972000	0.95
46	15	0	6	0.974000	0.95
47	15	0	11	0.976000	0.95
48	15	2	6	0.934000	0.95
49	15	2	11	0.934000	0.95
50	20	-2	6	0.936000	0.95
51	20	-2	11	0.938000	0.95
52	20	0	0	0.968000	0.95
53	20	0	6	0.970000	0.95
54	20	0	11	0.970000	0.95
55	20	2	6	0.928000	0.95
56	20	2	11	0.942000	0.95
57	50	-2	6	0.926000	0.95
58	50	-2	11	0.936000	0.95
59	50	0	0	0.940000	0.95
60	50	0	6	0.944000	0.95
61	50	0	11	0.946000	0.95
62	50	2	6	0.940000	0.95
63	50	2	11	0.942000	0.95
64	100	-2	6	0.942000	0.95
65	100	-2	11	0.934000	0.95
66	100	0	0	0.952000	0.95
67	100	0	6	0.958000	0.95
68	100	0	11	0.956000	0.95
69	100	2	6	0.940000	0.95
70	100	2	11	0.954000	0.95

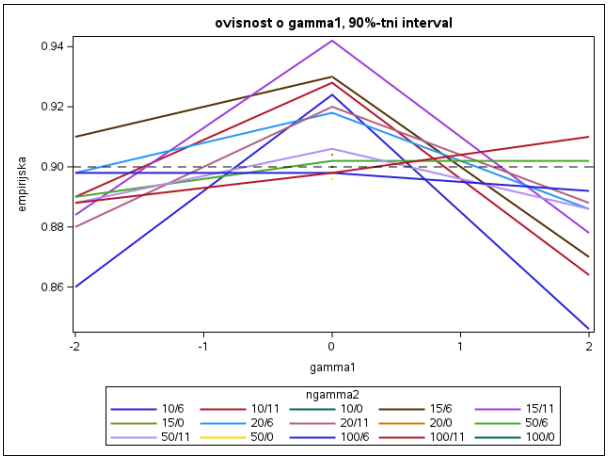
Slika 2: Tablica s rezultatima, nominalno 95%-p.i.



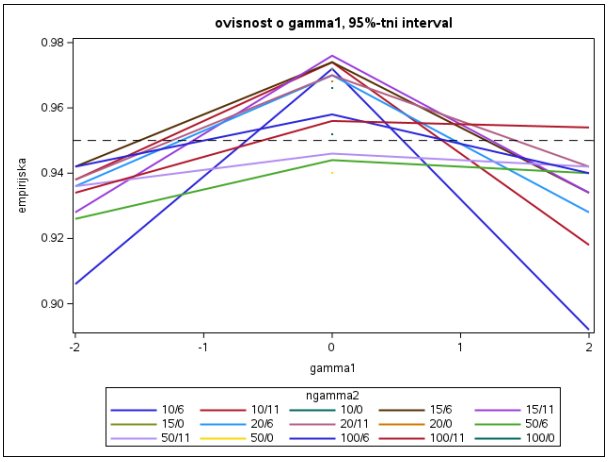
Slika 3: Ovisnost vjerojatnosti o n za fiksne γ_1, γ_2, 90%-p.i.



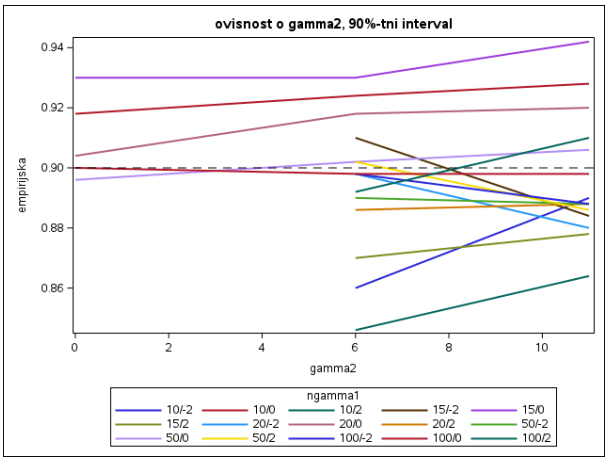
Slika 4: Ovisnost vjerojatnosti o n za fiksne γ_1, γ_2, 95%-p.i.



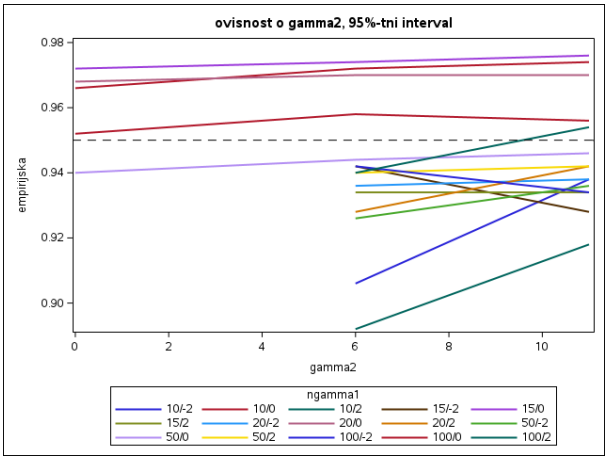
Slika 5: Ovisnost vjerojatnosti o γ_1 za fiksne n, γ_2, 90%-p.i.



Slika 6: Ovisnost vjerojatnosti o γ_1 za fiksne n, γ_2, 95%-p.i.



Slika 7: Ovisnost vjerojatnosti o γ_2 za fiksne n, γ_1, 90%-p.i.



Slika 8: Ovisnost vjerojatnosti o γ_2 za fiksne n, γ_1, 95%-p.i.

Projekt 5/2024.

Provedite MC studiju za ispitivanje utjecaja distribucije podataka na vjerojatnosti pokrivanja (engl. coverage probabilities) 95% i 90% t intervala pouzdanosti za sredinu populacije.

Navedeni eksperiment provedite za sve moguće kombinacije vrijednosti faktora:

n (veličina uzorka) = 10, 15, 20, 50, 100

skewness (koeficijent asimetrije) γ_1 = od -2 do 2 sa korakom 2 ,

kurtosis (koeficijent spljoštenosti) γ_2 = -3, 0 , 6, 11.

Vrijednosti sredine i standardne devijacije neka budu $\mu=0$ i $\sigma=5$.

Za svaku moguću kombinaciju faktora n, γ_1 i γ_2 izvedite 500 replikacija tako da se za svaku replikaciju generira n slučajnih brojeva sa sredinom $\mu=0$ i standardnom devijacijom $\sigma=5$.

Za svaki generirani uzorak izračunajte 95% i 90% interval pouzdanosti za sredinu (sa procedurom MEANS ili UNIVARIATE)

Za svaku kombinaciju n, γ_1 i γ_2 i za 95% i za 90% intervale pouzdanosti procijenite vjerojatnost pokrivanja tj. izračunajte proporciju uzoraka (od 500 uzoraka/replikacija) za koje se populacijska sredina $\mu=0$ nalazi unutar 95% i 90% intervala pouzdanosti (LCL < μ <UCL, gdje su LCL i UCL donja i gornja granica 95% odnosno 90% intervala pouzdanosti).

Kreirajte odgovarajuće tablice i grafikone. Usporedite procijenjene (stvarne) vjerojatnosti pokrivanja sa nominalnim vjerojatnostima (0.95 i 0.90) za pojedine n, γ_1 i γ_2.

Smije li se koristiti 95% (i 90%) interval pouzdanosti za n=10, kada je $\gamma_1 = 2, \gamma_2 = 11$?

NAPOMENA o mogućim kombinacijama vrijednosti koeficijenata asimetrije γ_1 i spljoštenosti γ_2:

Koeficijenti asimetrije γ_1 i spljoštenosti γ_2 moraju zadovoljavati slijedeće uvijete:

$$\gamma_1^2 - 2 \leq \gamma_2^2$$