Projektna naloga iz Statistike

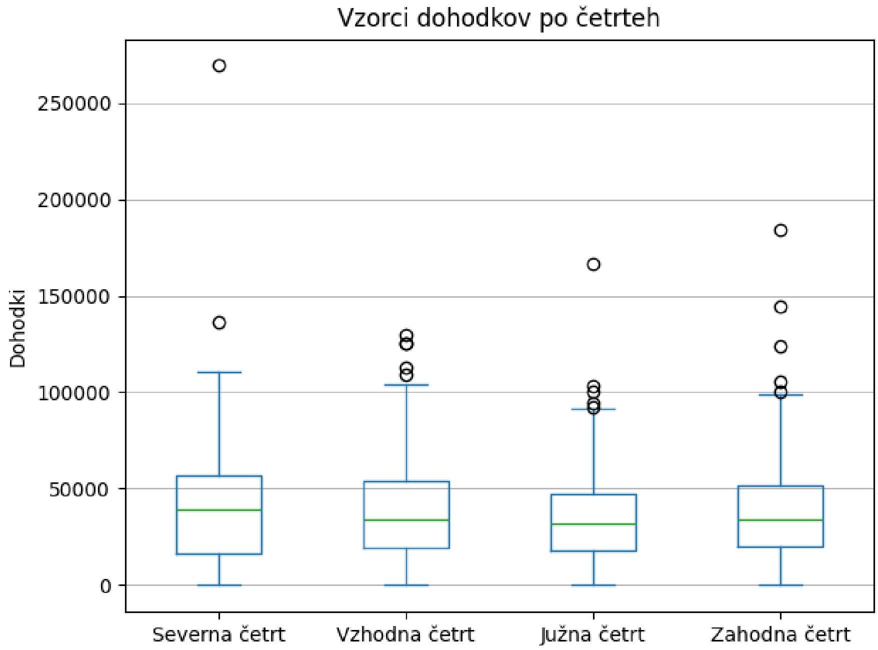
Jan Pantner (jan.pantner@gmail.com)

September 2024

1 Kibergrad

Preučujemo dohodke družin v mestu Kibergrad. Imamo informacije o 43.886 družinah, ki živijo v eni od štirih četrti: v severni četrti stanuje 10.149 družin, v vzhodni 10.390, v južni 13.457 in v zahodni 9.890. Pomagamo si s programom `kibergrad.py`.

Iz vsake četrti vzemimo enostavni slučajni vzorec velikosti 100. Dohodke lahko primerjamo s pomočjo vzporednih škatel z brki, ki so prikazane na sliki 1.

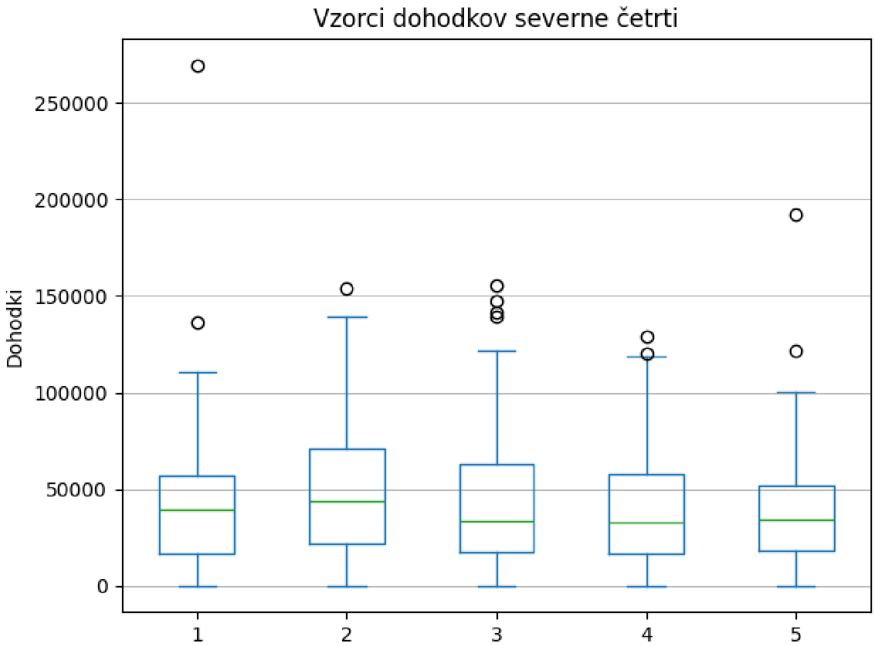


Slika 1: Škatle z brki za dohodke posamezne četrti.

Najprej opazimo, da je v severni četrti osamelec, ki močna izstopa. Prav tako so maksimum, tretji kvartil in mediana v severni četrti največji, vendar razlika ni dovolj velika, da bi lahko pri tako majhnem vzorcu sklepali, da so dohodki severne četrti najvišji. Zdi pa se, da so dohodki severne četrti malenkost višji kot dohodki južne četrti, pri kateri so vse prej omenjene vrednosti najmanjše.

Zdi se, da v splošnem četrt ne vpliva močno na velikost dohodka.

Sedaj vzemimo še štiri enostavne slučajne vzorce velikosti 100 iz severne četrti in poglejmo vzporedne škatle z brki za vzorce iz severne četrti. Prikazane so na sliki 2.



Slika 2: Škatle z brki dohodkov severne četrti.

Vse mediane so večje od 33.000, njihovo povprečje je približno 36.500. Zdi se, da lahko s precejšnjo sigurnostjo sklepamo, da je večina dohodkov višjih od 33.000.

V novih vzorcih ne opazimo tako ekstremnih osamelcev kot pri prvem vzorcu. Večina osamelcev ni zelo oddaljenih od maksimuma.

Za konec si poglejmo še varianco dohodka pojasnjeno s četrtmi in preostalo (rezidualno) varianco. Naj bo N velikost populacije, N_i velikost populacije i-te četrti in w_i velikostni delež i-te četrti. Naj bo μ povprečni dogodek in σ^2 varianca dohodka Kibergrada, μ_i in σ_i^2 pa zaporedoma povprečni dohodek in varianca dohodka i-te četrti. Pojasnjena in nepojasnjena varianca sta zaporedoma določeni s formulama

$$\sigma_B^2 := \sum_{i=1}^{4} w_i(\mu_i - \mu) \quad \text{in} \quad \sigma_W^2 := \sum_{i=1}^{4} w_i \sigma_i^2.$$

Izračunamo, da je

$$\sigma_B^2 = 9.252.923 \quad \text{in} \quad \sigma_W^2 = 1.017.226.451.$$

S četrti pojasnjeni standardni odklon je torej 3.042. Povprečni dohodki četrti, so enaki 45.759, 41.235, 37.473 in 42.158, torej je standardni odklon v primerjavi z njimi majhen. To se ujema z opažanjem, da četrt ne vpliva močno na velikost dohodka.

2 Lomljivost najlonskih palic

Na vzorcu 280 najlonskih palic preizkušamo njihovo lomljivost. Rezultati preizkusa so prikazani v tabeli 1. Pri analizi si pomagamo s programom `najlonske_palice.py`.

št. lomov	0	1	2	3	4	5
št. palic	157	69	35	17	1	1

Tabela 1: Rezultati preizkusa lomljivosti.

Privzemimo, da je število mest, na katerih se je palica zlomila porazdeljeno binomsko $\mathrm{Bin}(5, p)$ za določen neznan p. Privzemimo tudi, da so palice med seboj neodvisne.

Pri teh predpostavkah je logaritem verjetja podan z

$$l(p, x) = \log \prod_{i=1}^{n} \binom{5}{x_i} p^{x_i} (1-p)^{5-x_i}.$$

Izračunamo, da je maksimum dosežen pri približno $\hat{p} = 0{,}1421438163551369$. Ta vrednost nam predstavlja oceno p po metodi največjega verjetja.

Sedaj združimo zadnje tri vrednosti in naredimo posplošeni Pearsonov preizkus hi kvadrat. Preizkušamo ničelno domnevo, da je število mest, na katerih se je palica zlomila porazdeljeno binomsko $\mathrm{Bin}(5, p)$ proti alternativni domeni, da ima število lomov katero drugo porazdelitev.

Pearsonova testna statistika je

$$X^2 = \sum_{i=0}^{2} \frac{(x_i - n f_i(\hat{p}))^2}{n f_i(\hat{p})},$$

kjer je $x_0 = 157$, $x_1 = 69$, $x_2 = 35$, $x_3 = 19$ in

$$f_i(\hat{p}) = \begin{cases} \binom{5}{x_i} \hat{p}^{x_j} (1-\hat{p})^{5-x_i}, & i \in \{0,1,2\} \\ \sum_{j=3}^{5} \binom{5}{x_j} \hat{p}^{x_j} (1-\hat{p})^{5-x_j}, & i = 3 \end{cases}.$$

Izračunamo, da je $X^2 > 44$. Iz tabela kvantilov porazdelitve hi kvadrat sledi, da ničelno domnevo zavrnemo tako pri stopnji tveganja $\alpha = 0{,}01$ kot tudi pri $\alpha = 0{,}05$.