Projektna naloga pri predmetu Statistika

Beno Učakar
Profesor: doc. dr. Martin Raič

1. naloga

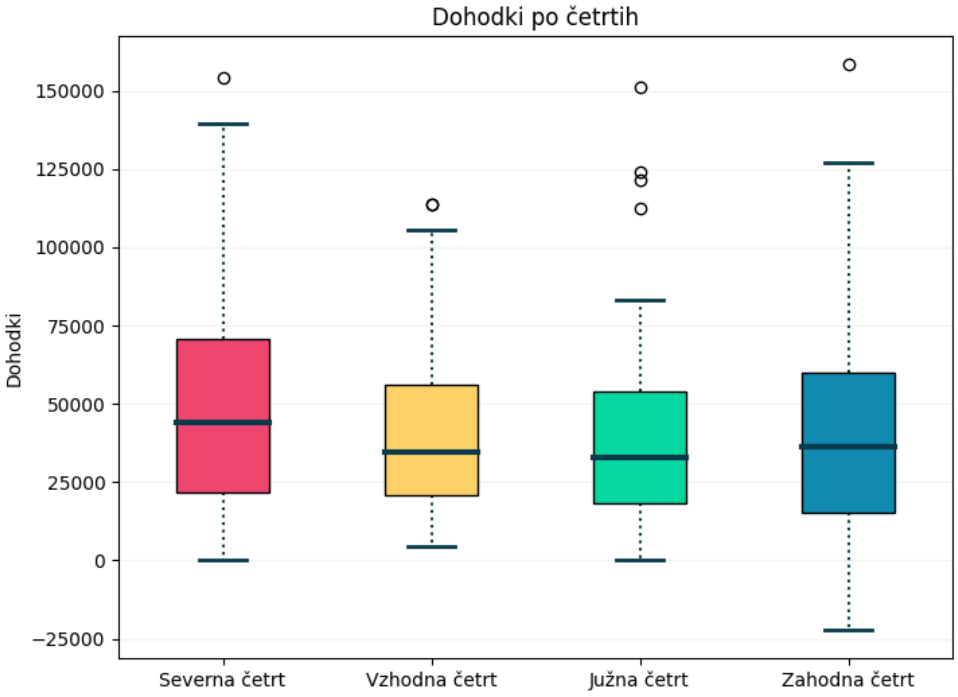
Nalogo rešujemo s pomočjo programa **naloga1.py**. Ta generira škatle z brki in za zadnji del naloge vrne:

```
Izhod
```

Preučevali bomo skupni dohodek družin v mestu Kibergrad. Imamo informacije o 43.886 družinah, ki so v enem od štirih četrti. Število družin v severni, vzhodni, južni oziroma zahodni četrti je 10.149, 10.390, 13.457 oziroma 9.890.

Primer (a)

Iz vsake četrti izberemo slučajni vzorec velikosti 100. Na podlagi teh vzorcev narišemo škatle z brki za dohodke po četrtih, ki so prikazane na sliki 1.



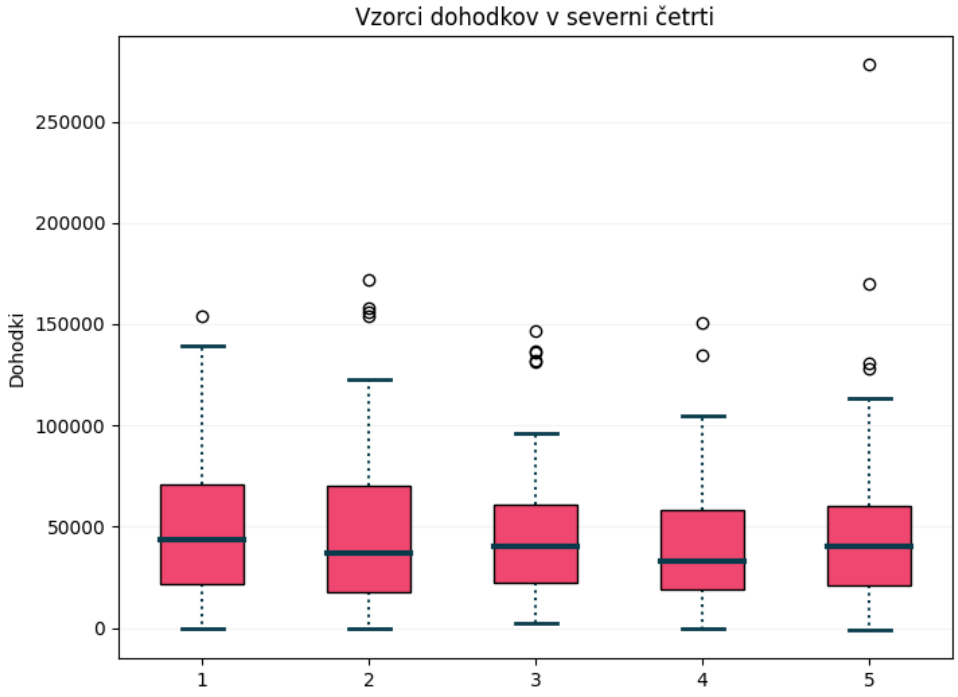
Slika 1: Škatle z brki za dohodke po četrtih.

Najprej naredimo nekaj splošnih opazk. Opazimo, da je variacija dohodkov znotraj severne in zahodne četrti nekoliko višja kot v vzhodni in južni četrti. Ker so prvi, drugi in tretji kvartil ter maksimum v severni četriti izmed vseh četrti največji, sklepamo, da je v povprečju dohodki v Severni četrti malenkost višji od ostalih. V vzhodni in južni četrti je porazdelitev nekoliko nagnjena k večjim dohodkom. Južna četrt ima največ osamelcev, dohodki pa so izmed vseh četrti najmanj razpršeni.

Na podlagi opaženega, sklepamo, da so dohodki v severni četrti nekoliko višji kot v ostalih. Pravtako pa se zdi, da je varianca povprečnega dohodka med četrtmi relativno majhna. Glede na to da smo iz vsake četrti izbrali zgolj 100 vzorcev, populacija četrti pa je v povprečju 10.000, ti podatki niso nujno reprezentativni. Zato s tem sklepom postopamo previdno.

Primer (b)

Iz severne četrti vzamemo še 4 vzorce velikosti 100. Tudi za te vzorce narišemo škatle z brki prikazane na sliki 2.



Slika 2: Škatle z brki za dohodke v severni četrti.

Mediane vseh škatel ležijo nad 40.000, tako da sklepamo da večina dohodkov znaša več kot to. Dohodki osamelcev znašaja v povprečju 150.000. Nasploh opazimo nekoliko več variacije med premožnejšimi prebivalci severne četriti.

Primer (c)

Naj bo N velikost populacije Kibergrada, N_i velikost populacije i-te četrti in $w_i = N_i/N$ velikostni deleži četrti. Nadaljnje naj bo μ_i povprečni dohodek in σ_i^2 varianca dohodka i-te četrti, μ povprečni dohodek in σ^2 varianca dohodka celotne populacije ter σ_p^2 in σ_n^2 pojasnjena in nepojasnjena varianca celotne populacije. Pojasnjeno in nepojasnjeno varianco pri stratificiranem vzorčenju lahko izrazimo kot

$$\sigma_p^2 = \sum_{i=1}^{4} w_i \mu_i^2 - \mu^2 \qquad \sigma_n^2 = \sum_{i=1}^{4} w_i \sigma_i^2.$$

Programa **naloga1.py** vrne, da pojasnjena varianca znaša, nepojasnjena varianca pa. Nizka pojasnjena varianca potrjuje hipotezo, da je razlika povprečnega dohodka družine med četrtmi majhna.

2. naloga

Primer (a)

Najprej uvedimo nekaj oznak. Če je $k \in \{1, \ldots, 12\}$ število skokov ptic, naj bo S_k frekvenca tega opažanja. Podatke z novimi oznakami predstavimo v spodnji tabeli.

k	1	2	3	4	5	6	7	8	9	10	11	12
S_k	48	31	20	9	6	5	4	2	1	1	2	1

Tabela 1: Frekvence števila skokov

Skupno število opaženih skokov označimo z S, število vseh opažanj pa z N. Velja

$$N = \sum_{k=1}^{12} S_k \qquad S = \sum_{k=1}^{12} kS_k.$$

V našem primeru znaša $N = 130$ in $S = 363$.

Želimo poiskati geometrijsko porazdelitev, ki se najbolje prilega tem podatkom. Če je število skokov pri posameznem opažanju slučajna spremenljivka $K \sim \text{Geom}(p)$, v resnici iščemo cenilko za parameter p. To znamo narediti na vsaj dva načina.

Prvi način: Postopamo po metodi momentov. Spomnimo se, da pričakovana vrednost geometrijske porazdelitve Geom(p) znaša $\frac{1}{p}$. Zato velja

$$p = \frac{1}{E(K)}.$$

Po metodi momentov E(K) ocenimo s prvim momentom opaženih vrednosti

$$\frac{1}{N} \sum_{k=1}^{12} kS_k = \frac{S}{N},$$

kar nam da cenilko

$$\hat{p} = \frac{N}{S}.$$

Drugi način: Postopamo po metodi največjega verjetja. Verjetnostna funkcija geometrijske porazdelitve Geom(p) je $\text{P}(K = k) = p(1-p)^{k-1}$. Verjetje lahko torej izrazimo kot

$$\text{L}(p \mid S1, \ldots, S_{12}) = p^N (1-p)^{S-N}.$$

Ko logaritmiramo, dobimo

$$\text{l}(p \mid S1, \ldots, S_{12}) = N \ln(\frac{p}{1-p}) + S \ln(1-p).$$

Če parcialno odvajamo po p in malo računamo, ponovno pridemo do cenilke

$$\hat{p} = \frac{N}{S}.$$

V obeh primerih pridemo do iste cenilke. Ta v našem primeru znaša

$$\hat{p} = \frac{130}{363} \approx 0.358.$$

Teorija metode momentov in metode največjega verjetja nam zagotovita, da je ta izbira smiselna. Iskana geometrijska porazdelitev je $\text{Geom}(\hat{p})$.

Primer (b)

Ob predpostavki, da je $K \sim \text{Geom}(\hat{p})$, poračunamo verjetnosti p_k, da pri enem opažanju pride do k skokov. Pričakovano vrednosti frekvenc določimo kot

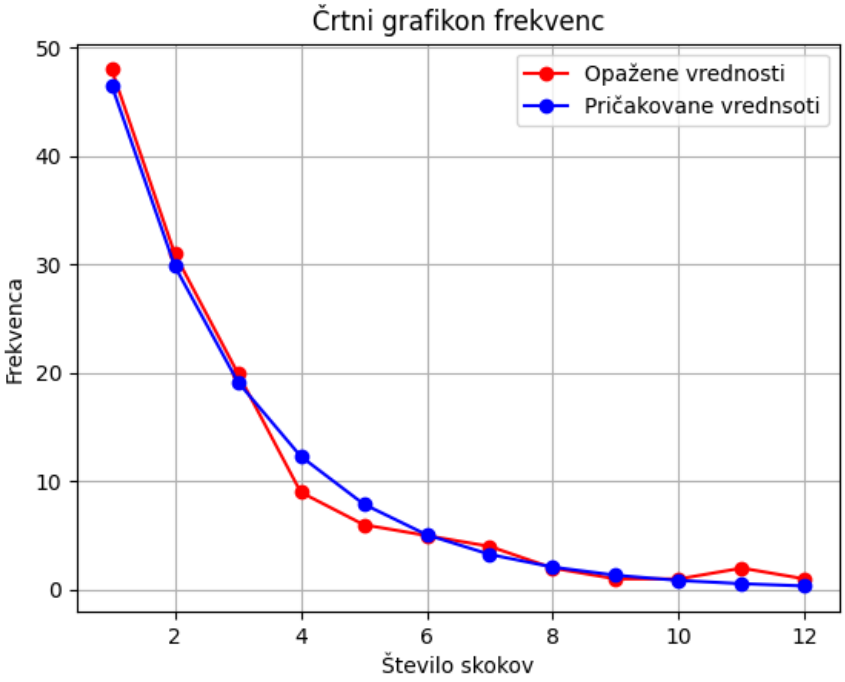
$$\hat{S}_k = p_k N.$$

Dobimo spodnjo tabelo.

k	1	2	3	4	5	6
p_k	0.3580	0.2298	0.1476	0.0947	0.0608	0.0390
$\hat{S}_k$	46.54	29.87	19.19	12.31	7.90	5.07
k	7	8	9	10	11	12
p_k	0.0251	0.0161	0.0103	0.0066	0.0043	0.0027
$\hat{S}_k$	3.26	2.09	1.34	0.86	0.56	0.35

Tabela 2: Pričakovane frekvence skokov.

Te podatke združimo v spodnji črtni grafikon.



Slika 3: Črtni grafikon opaženih in pričakovanih prekvenc.

Primer (c)

Izračunati moramo pričakovano vrednost naše cenilke. Naj bo $K_i \sim \text{Geom}(p)$ število skokov pri i-tem opažanju. Opazimo, da je

$$S = \sum_{i=1}^{N} K_i.$$

Če predpostavimo, da so spremenljivke $K_1, K_2, \ldots, K_N$ med sabo neodvisne, je slučajna spremenljivka S porazdeljena negativno binomsko $\text{NegBin}(N, p)$. Računamo.

$$E(\hat{p}) = E(\frac{N}{S}) = \sum_{k=N}^{\infty} \frac{N}{k} \binom{k-1}{N-1} p^N (1-p)^{N-k}$$