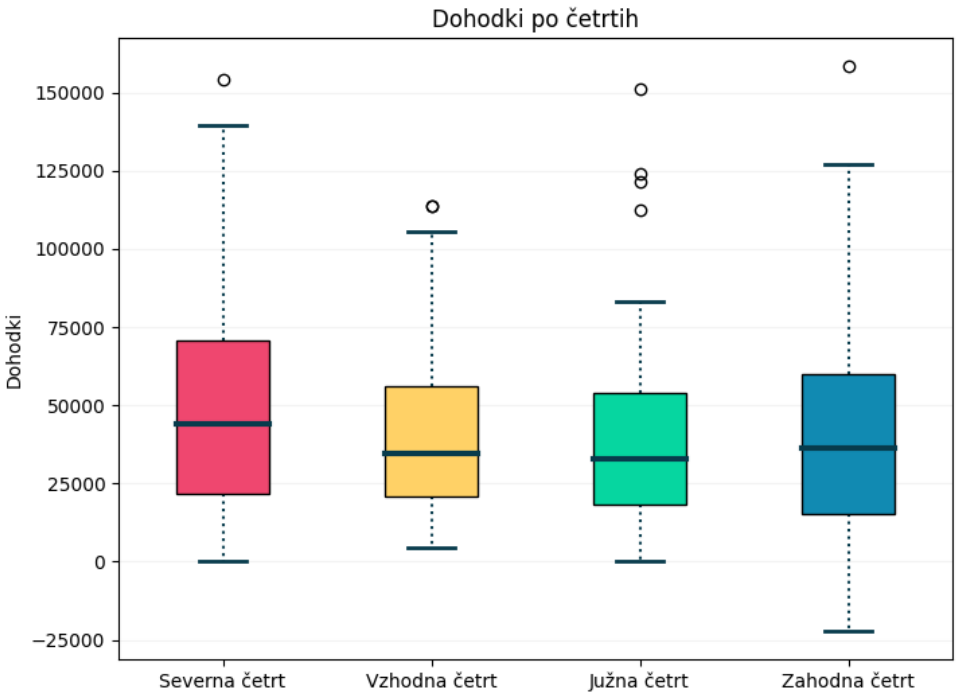
Projektna naloga pri predmetu Statistika

Beno Učakar
Profesor: doc. dr. Martin Raič

1. naloga

Primer (a)

V programu Python najprej uvozimo podatke in ločimo četrti. Nato iz vsake četrti izberemo slučajni vzorec velikosti 100. Narišemo vzporedne škatle z brki za dohodke po četrtih.



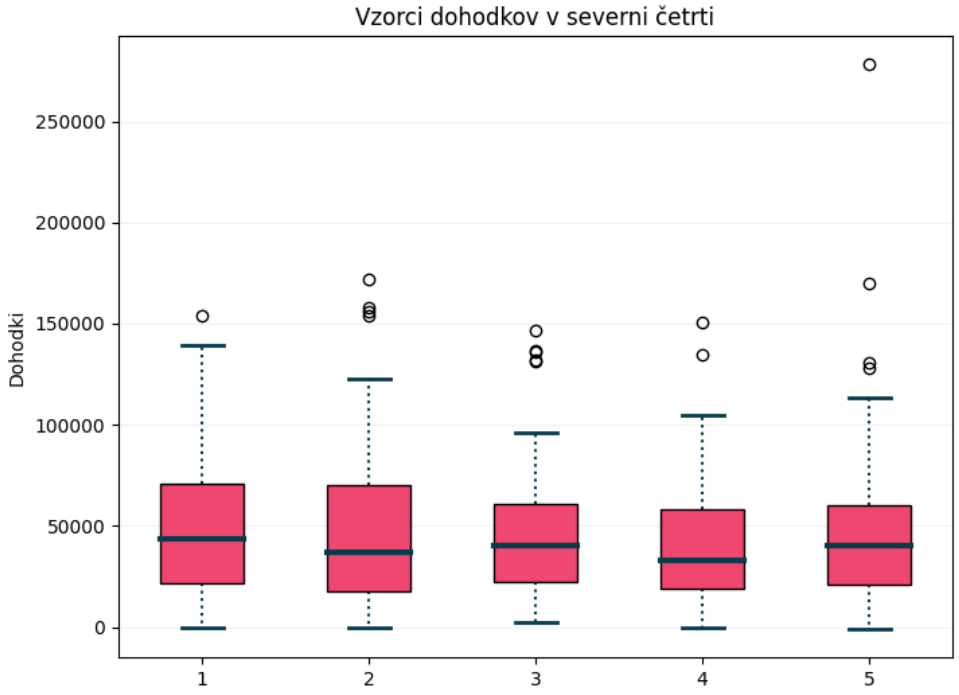
Slika 1: Škatle z brki za dohodke po četrtih.

Opazimo, da je variacija dohodkov v severni in zahodni četrti nekoliko višja kot v vzhodni in južni četrti. Razberemo lahko, da je v povprečju do-

hodki v Severni četrti malenkost višji na podlagi največjega tretjega kvartila. Južna četrt ima največ osamelcev, dohodki večine pa so sicer izmed vseh četrti najnižji.

Glede na to da smo iz vsake četrti izbrali zgol 100 vzorcev, populacija četrti pa je v povprečju 10.000, ti podatki niso ravno reprezentativni. Tu ne moremo sklepati prav veliko.

Primer (b)



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

Primer (c)

Na vajah smo izpeljali enačbe za pojasnjeno in nepojasnjeno varianco pri stratificiranem vzorčenju. Naj bo N velikost populacije, N_i velikosti posameznih četrti za $k \in \{1, 2, 3, 4\}$, $w_i = N_i/N$ velikostni deleži četrti, n_i velikost vzorca i-te četrti, μ povprečni dohodek populacije, μ_i povprečni dohodek i-te četrti, σ^2 varianca dohodka populacije in σ_i^2 varianca dohodka i-te četrti. Populacijsko povprečje lahko izrazimo kot

$$\mu = w_1\mu_1 + w_2\mu_2 + w_3\mu_3 + w_4\mu_4.$$

Označimo pojasnjeno varianco z σ_p^2 in nepojasnjeno varianco σ_n^2. Potem velja

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

Naj bo X_{ij} j-ti vzorec i-te četrti, $\bar{X}$ povprečje celotnega vzorca $\bar{X}_i$ in $\bar{X_i^2}$ prvi in drugi vzorčni moment i-te četrti ter $S_i = \sum_{i=1}^{n_i} (X_{ij} - \bar{X}_i)^2$. Nepristrasnki cenilki za σ_p^2 in σ_n^2 sta

$$\hat{\sigma_p^2} = \sum_{i=1}^{4} w_i \frac{N_i - 1}{N_i} \frac{S_i}{n_i - 1}$$

$$\hat{\sigma_p^2} = \sum_{i=1}^{4} w_i \left[(\bar{X_i^2} - \bar{X})^2 + \frac{N_i - 1}{N_i} \frac{S_i}{n_i - 1} - (1 - w_i) \frac{N_i - n_i}{N_i} \frac{S_i}{n_i (n_i - 1)} \right].$$

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

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

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

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

kar nam da cenilko

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

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

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

Ko logaritmiramo, dobimo

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

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 Geom($\hat{p}$).

Primer (b)

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

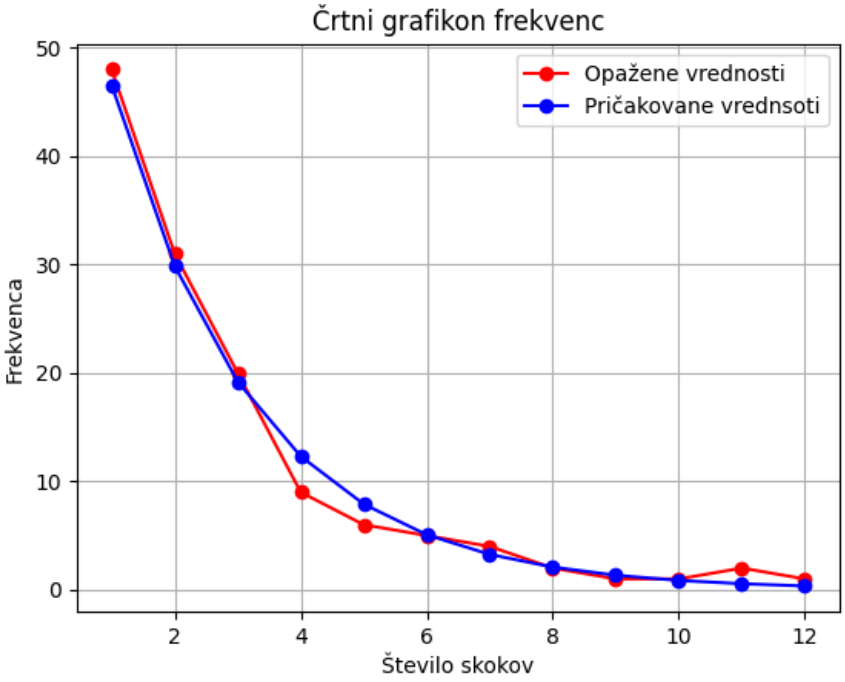
$$\hat{S}_k = p_k N.$$

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.

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