Projektna naloga pri predmetu Statistika

Beno Učakar
Profesor: doc. dr. Martin Raič

1. naloga

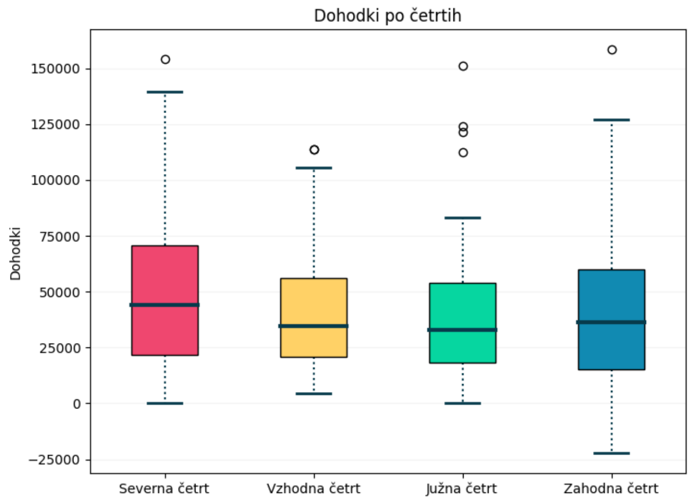
Nalogo rešujemo s pomočjo programa **naloga1.py**. Ta generira škatle z brki in za zadnji del naloge vrne:

```
S četrtmi pojasnjena varianca dohodka družin Kibergrada
znaša 9252923, residualna varianca pa znaša 1017132747.
Pojasnjeni standardni odklon dohodka med četrtmi znaša 3042.
Povprečni dohodki znaša 45759 v severni, 41235 v vzhodni,
37473 v južni in 42158 v zahodni četrti.
```

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.



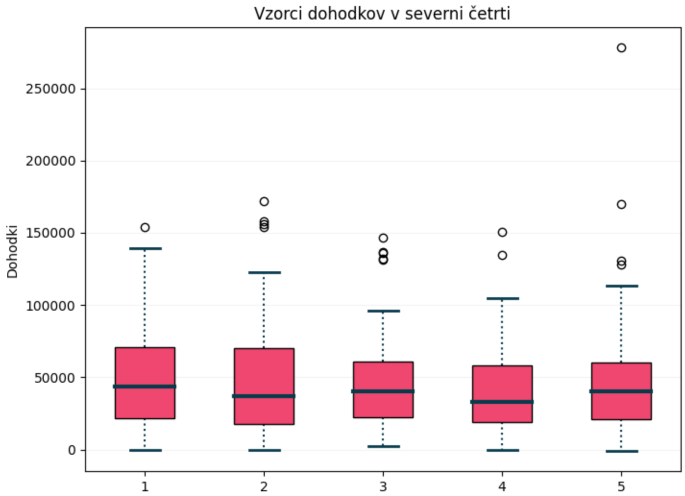
Slika 1: Škatle z brki za dohodke po četrtih.

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 $w_i = \frac{N_i}{N}$ 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 9.252.923, nepojasnjena varianca pa znaša 1.017.132.747. Pojasnjeni standardni odklon dohodka med četrtmi znaša 3.042, kar je malo v primerjavi s povprečnimi dohodki četrti, ki znšajo 45.759 v severni, 41.235 v vzhodni, 37.473 v južni in 42.158 v zahodni četrti. To 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 $K \sim \text{Geom}(p)$, 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 $\text{Geom}(p)$ znaša $\frac{1}{p}$. Zato velja

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

Po metodi momentov $\text{E}(K)$ 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 $\text{Geom}(p)$ je $\text{P}(K = k) = p(1-p)^{k-1}$. Verjetje lahko torej izrazimo kot

$$\text{L}(p \mid S_1, \ldots, S_{12}) = p^N (1-p)^{S-N}.$$

Ko logaritmiramo, dobimo

$$\text{l}(p \mid S_1, \ldots, S_{12}) = N \ln(\frac{p}{1-p}) + S \ln(1-p).$$

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 $\mathrm{Geom}(\hat{p})$.

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

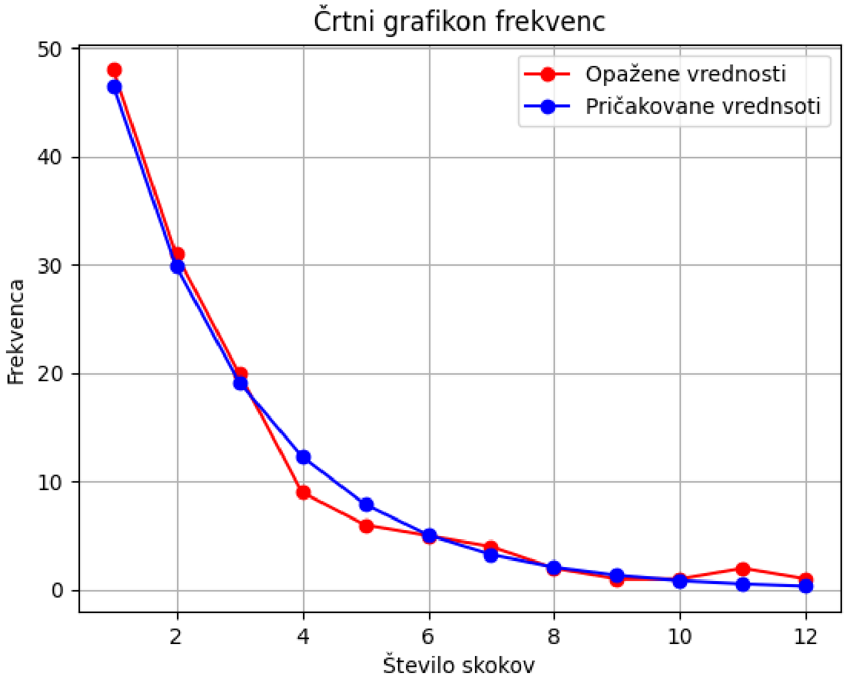
$$\hat{S}_k = p_k N.$$

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 $K_i \sim \mathrm{Geom}(p)$ 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 $\mathrm{NegBin}(N, p)$. Računamo.

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

3. naloga

Nalogo rešujemo s pomočjo programa **naloga3.py**. Ta generira histogram in vrne naslednje podatke:

```
Studentova statistika znaša 19.29739 in ima
```

```
107 prostorskih stopenj.
Širina stolpcev po modificiranem Freedman-Diaconisovem
pravilu znaša 20.499.
Studentova statistika znaša 0.09837 in ima
(2, 43) prostorskih stopenj.
```

Naj bo P_i^1 prvi in P_i^2 drugi izmerjen pulz i-tega študenta. Za vsakega študenta izračunamo spremembo pulza $\Delta P_i = P_i^2 - P_i^1$. To statistiko bomo obravnavali v nadaljevanju.

Primer (a)

Študente razdelimo v dve skupini glede na to ali so bili med meritvama pulzov deležni obremenitve ali ne. Za $i = 1, \ldots, n$ spremembo pulza i-tega študenta, ki je bil deležen obremenitve, označimo z X_i. Podobno za $j = 1, \ldots, m$ spremembo pulza j-tega študenta, ki ni bil deležen obremenitve, označimo z Y_j. Predpostavimo, da so razlike pulzov X_i porazdeljene normalno s porazdelitvijo $\mathrm{N}(\mu_X, \sigma^2)$, razlike pulzov Y_j pa normalno s porazdelitvijo $\mathrm{N}(\mu_Y, \sigma^2)$.

Testiramo ničelno hipotezo

$$H_0 : \mu_X = \mu_Y$$

proti enostranski alternativni hipotezi

$$H_1 : \mu_X > \mu_Y.$$

V ta namen opravimo T-test na testni statistiki

$$T = \frac{\bar{X} - \bar{Y}}{s_p \sqrt{\frac{1}{n} + \frac{1}{m}}}.$$

Program **naloga3.py** vrne, da statistika T znaša 19,29739 in ima 107 prostorskih stopenj. Iz tabel razberemo, da znašata

$$F^{-1}_{\text{Student}(60)}(0.05) = 1{,}671 \quad \text{in} \quad F^{-1}_{\text{Student}(60)}(0.01) = 2{,}390.$$

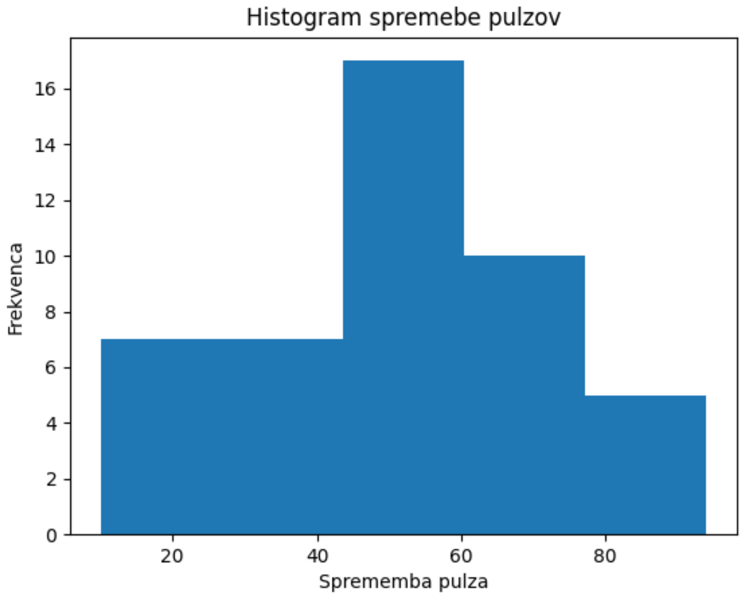
Ker funkcija $F^{-1}_{\text{Student}(k)}$ pada v odvisnosti od parametra k, velja

$$T > F^{-1}_{\text{Student}(107)}(0.05) \quad \text{in} \quad T > F^{-1}_{\text{Student}(107)}(0.01).$$

Ničelno hipotezo zavrnemo pri stopnji tveganja 0,05 in 0,01. Zanesljivo lahko trdimo, da obremenitev vpliva na spremembo pulza.

Primer (b)

Za spremembe pulzov študentov, ki so bili določeni za obremenitev, narišemo histogram, ki je prikazan na sliki 4. Širine stolpcov so določene v skladu z modificiranem Freedman-Diaconisovem pravilu in znaša 20,499.



Slika 4: Histogram spremembe pulzov pri študentih, ki so bili določeni za obremenitev.

Po predpostavki, naj bi bile spremembe pulzov pri študentih, ki so bili določeni za obremenitev, normalno porazdeljene. Iz zgornjega histograma pa je razvidno, da je nekoliko več študentov imelo spremembo pulza pod 30, kot pa bi pričakovali od normalne porazdelitve. Zdi se, da je res nekaj študentov goljufalo.

Primer (c)

Vseh N študente, ki so bili deležni obremenitve, razdelimo v $k = 3$ skupine, glede na njihovo vadbo. Z X_{ij} označimo spremembo pulza j-tega študenta i-te skupine. Predpostavimo model

$$X_{ij} = \mu + \alpha_i + \epsilon_{ij},$$

kjer je μ pričakovana vrednost vseh meritev, α_i odstopanje od pričakovane vresnosti i-te skupine in ϵ_{ij} šum, za katere pa predpostavimo, da so med sabo neodvisni in porazdeljeni normalno $\mathrm{N}(0, \sigma^2)$.

Testiramo ničelno hipotezo

$$H_0 : \alpha_1 = \alpha_2 = \alpha_3 = 0$$

proti alternativni hipotezi

$$H_1 : \text{Niso vsi } \alpha_i = 0.$$

V ta namen opravimo ANOVA F-test na statistiki

$$F = \frac{\frac{SS_B}{k-1}}{\frac{SS_W}{N-k}},$$

kjer je

$$SS_B = \sum_{i=1}^{k} N_i(\bar{X}_i - \bar{X})^2 \quad \text{in} \quad \sum_{i=1}^{k}\sum_{j=1}^{N_i}(X_{ij} - \bar{X}_i)^2.$$

Tu smo z N_i označili velikost i-te skupine, z $\bar{X}_i$ vzorčno povprečje i-te skupine in z $\bar{X}$ vzorčno povprečje vseh meritev. Program **naloga3.py** vrne, da statistika F znaša $0,09837$ in ima $(2,43)$ prostorskih stopenj. Iz tabel razberemo, da znašata

$$F^{-1}_{\text{Fisher}(2,60)}(0.05) = 3{,}15 \quad \text{in} \quad F^{-1}_{\text{Fisher}(2,60)}(0.01) = 4{,}98.$$

Ker funkcija $F^{-1}_{\text{Fisher}(2,k)}$ pada v odvisnosti od parametra k, velja

$$F < F^{-1}_{\text{Fisher}(2,43)}(0.05) \quad \text{in} \quad F < F^{-1}_{\text{Fisher}(2,43)}(0.01).$$

Tako ne moremo niti pri stopnji tveganja 0,05 niti pri 0,01 zavrniti ničelne hipoteze. Nimamo dovolj podatkov, da bi lahko trdili, da vadba vpliva na spremembo pulza.