Projekt statistika

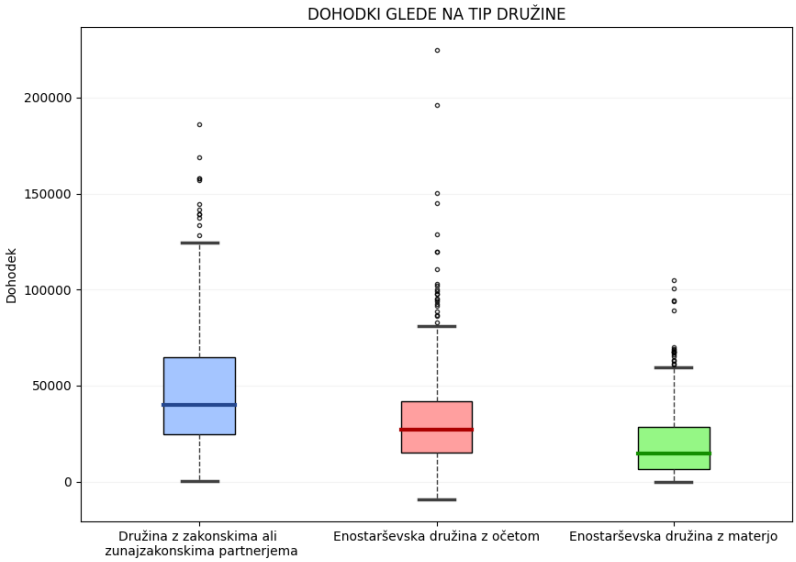
Projekt pri predmetu Statistika

Matej Novoselec

30. junij 2023

1 Kibergrad

Podani so nam podatki o dohodkih 43886 družin iz mesta Kibergrad. Podatke o dohodkih bomo analizirali v odvisnosti od tipa družin, ki je posredovala podatke. Pri obdelavi podatkov si bomo pomagali z datoteko *statistika_naloga_1.py*, ki je namenjena predvsem izrisu grafov, ter numeričnim izračunom željenih vrednosti, prek danih podatkov.
Družine razdelimo v 3 tipe: tip 1 - družina z zakonskima ali zunajzakonskima partnerjema, tip 2 - enostarševska družina z očetom, tip 3 - enostarševska družina z materjo. Imamo 33403 podatkov od družin tipa 1, 2054 podatkov od družin tipa 2 in 8429 podatkov od družin tipa 3. Za vsak tip družine izberemo enostaven slučajen vzorec velikosti 500 in narišemo pripadajočo škatlo z brki.

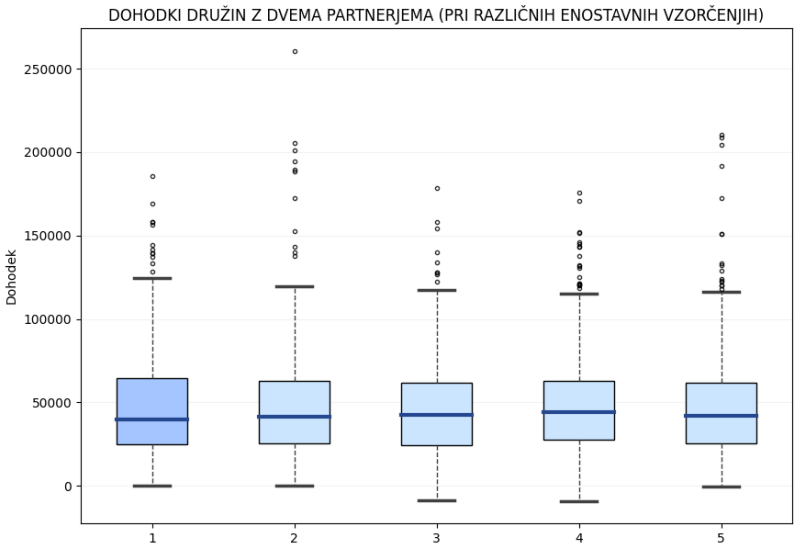


Slika 1: Škatle z brki za dohodke različnih tipov družin

Hitro vidimo, da so vrednosti prvega, drugega in tretjega kvantila pri družinah tipa 1 (družinah z zakonskima ali zunajzakonskima partnerjema) višji od prvega, drugega in tretjega kvantila pri družinah tipa 2 in tipa 3. Iz tega lahko sklepamo, da so dohodki omenjenega tipa družin večji od obeh ostalih tipov. Na podoben način lahko sklepamo tudi, da so dohodki v enostarševski družini z očetom v povprečju nekoliko višji od dohodkov enostarševskih družin z materjo. Opazimo tudi, da je varianca dohodkov pri družinah tipa 1 večja od družin preostalih tipov. Osamelcev opazimo največ pri družinah tipa 2, pri družinah tipa 3 pa so podatki najmanj razpršeni. Vredno je omeniti tudi, da je pri družinah tipa 1 in tipa 3 porazdelitev dohodkov nekoliko nagnjena k dohodkov višje vre-

dnosti.
Povzeli bi lahko, da nekoliko izstopajo podatki o družinah tipa 1, ki imajo povprečno najvišji dohodek, a imajo obenem tudi veliko varianco, njihova porazdelitev pa je nagnjena k višjim vrednostim. Vseeno moramo biti nekoliko pazljivi, saj je vzorec velikosti 500 za 33403 podatkov relativno majhen.

Sedaj se osredotočimo na družine tipa 1 in iz podatkov izberimo še štiri enostavne slučajne vzorce. Opažanja so prikazana z vzporednimi škatlami z brki, pri čemer temnejše obarvana škatla z brki pripada podatkom, ki so bili izbrani kot enostavni slučajni vzorec za družine tega tipa iz prejne podnaloge.



Slika 2: Škatle z brki za dohodke družin z dvema partnerjema

Opazimo, da je bil zgoraj izbrani vzorec, kljub temu, da je bil dokaj majhen glede na število vseh podatkov, v veliki meri (vsaj iz vidika ostalih štirih vzorcev) representativen. Vidno je namreč, da so vrednosti prvega in tretjega kvartila pri vzorcih dokaj usklajeni. Tudi iz vidika števila osamelcev je prvotno izbran vzorec representativen. Nekoliko se od representativnosti odmaknemo le pri vrednosti drugega kvantila, ki je pri našem vzorcu nekoliko nižje kot pri ostalih, posledično pa so dohodki nekoliko nagnjeni k višjim vrednostim.

Zadnje podnaloge se za začetek lotimo nekoliko bolj teoretično, ter šele nato vstavimo podatke za izračun željenih vrednosti. Z n označimo število vseh numeričnim podatkov, t.j. skupni numerus je enak n. Podatki so razdeljeni v skupine (glede na tip družine), z n_i označimo število podatkov za skupino s

tipom družine i (i seveda tu 1, 2 ali 3). Definirajmo še uteži za posamezno skupino, kot $w_i = n_i/n$.
Pričakovano vrednost za celoten Kibergrad lahko izrazimo prek pričakovanih vrednosti za posamezne skupine (označimo jih z μ_i) kot:

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

Če z σ_i^2 označimo varianco znotraj i-te skupine, lahko na podoben način varianco dohodka za celoten Kibergrad izrazimo kot:

$$\sigma^2 = \sigma_P^2 + \sigma_N^2, \quad \text{kjer: } \sigma_P^2 = \sum_{i=1}^{3} w_i(\mu_i - \mu)^2 \quad \text{in} \quad \sigma_N^2 = \sum_{i=1}^{3} w_i\sigma_i^2.$$

Izpeljava je navedena v [2]. V zgornjem zapisu σ_P^2 označuje s tipom družine pojasnjeno varianco, σ_N^2 pa nepojasnjeno varianco. Naloga nas sprašuje po obeh. Numerično v *python* datoteki poračunamo približke za pričakovane vrednosti in variance za posamezne skupini, ter po zgoraj zapisanih formulah dobimo, da s tipom družine pojasnjena varianca znaša 113781161.94, nepojasnjena varianca pa 912604507.95. Omenimo še, da delež pojasnjene variance ($\eta^2 = \frac{\sigma_P^2}{\sigma^2}$) znaša 0.11085, pojasnjeni standardni odklon pa 10666.83.

Oglejmo si novo pridobljene podatke v luči preteklih opažanj, predvsem iz vidika povprečnih dohodkov različnim tipov družin. Že v komentarju prve podnaloge smo omenili, da so med povprečnimi dohodki tipov družin opazne razlike. Sedaj opazimo, da je tudi vrednost pojasnjenega standardnega odklona visoka glede na povprečne dohodke posameznih tipov družin (te znašajo 47187.48 za tip 1, 31637.36 za tip 2 in 20508.19 za tip 3), kar dodatno nakazuje na razlike med dohodki posameznih tipov družin. Vrednost deleža pojasnjene variance le dodatno potrdi že zapisano.

2 Naloga 2

Vredno je zapisati nekoliko bolj matematično interpretacijo problema/navodila. Podatki pripadajo trem različnim eksperimentom, ki bodo določali končne vrednosti in rezultate, a so problemi, ki jih podajajo, teoretično iste narave. Za podan problem sedaj razvijemo teoretični pristop in se lotimo reševanja podnaloge a) in b).
Imamo n neodvisnih, enako porazdeljenih slučajnih spremenljivk, označimo jih z $R_1,\ R_2,\ \ldots,\ R_n$. Porazdeljene naj bodo Rayleighovo, t.j. z gostoto:

$$f_{R_i}(r_i \mid \theta) = \begin{cases} \frac{r_i}{\theta^2} \exp\left(-\frac{r_i^2}{2\theta^2}\right) & ;\ \ r > 0 \\ 0 & ;\ \text{sicer} \end{cases} .$$

Zaradi predpostavljene neodvisnosti, je potem $(R_1,\ R_2,\ \ldots,\ R_n)$ porazdeljen z gostoto:

$$\prod_{i=1}^{n} f_{R_i}(r_i \mid \theta) = \left(\frac{1}{\theta^{2n}} \prod_{i=1}^{n} r_i\right) \exp\left(-\frac{1}{2\theta^2} \sum_{i=1}^{n} r_i^2\right); \quad r_i > 0.$$

Lotimo se podnaloge a). Velja:

$$L(\theta \mid (r_1,\ r_2,\ \ldots,\ r_n)) = \prod_{i=1}^{n} L_i(\theta \mid r_i) = \prod_{i=1}^{n} f_{R_i}(r_i \mid \theta)$$

in

$$l(\theta \mid (r_1,\ r_2,\ \ldots,\ r_n)) = \ln L(\theta \mid (r_1,\ r_2,\ \ldots,\ r_n)) = \sum_{i=1}^{n} \ln\big(f_{R_i}(r_i \mid \theta)\big),$$

ter zato:

$$l(\theta \mid (r_1,\ r_2,\ \ldots,\ r_n)) = -2n\ln(\theta) + \sum_{i=1}^{n} \ln(r_i) - \frac{1}{2\theta^2} \sum_{i=1}^{n} r_i^2 \ .$$

Iščemo cenilko za θ po metodi največjega verjetja, zato si ogledamo enakost:

$$0 = \frac{\partial l(\theta \mid (r_1,\ \ldots,\ r_n))}{\partial \theta} = -\frac{2n}{\theta} + \frac{1}{\theta^3} \sum_{i=1}^{n} r_i^2.$$

Za cenilko po metodi največjega verjetja tako dobimo:

$$\hat{\theta}_{MNV} = \sqrt{\frac{1}{2n} \sum_{i=1}^{n} r_i^2} \ .$$

Za rešitev podnaloge b) si oglejmo pričakovano vrednost (Rayleighove) slučajne spremenljivke R:

$$\mathbb{E}(R) = \int_0^\infty r\ f(r \mid \theta)\ dr = \int_0^\infty \frac{r^2}{\theta^2}\ \exp\left(-\frac{r^2}{2\theta^2}\right)\ dr.$$

Uvedemo $\tau = \frac{r^2}{2\theta^2}$ in dobimo

$$\mathbb{E}(R) = \int_0^\infty \sqrt{2\tau}\ \theta\ e^{-\tau}\ d\tau = \theta\sqrt{2}\int_0^\infty \tau^{1/2}\ e^{-\tau}\ d\tau = \theta\sqrt{2}\ \Gamma(3/2) = \theta\sqrt{\frac{\pi}{2}}.$$

Cenilka po metodi momentov je zato podano s predpisom:

$$\hat{\theta}_{MM} = \overline{R}\sqrt{\frac{2}{\pi}}.$$

Iz izpeljave cenilke, dobljene po metodi momentov, je jasno vidno, da je cenilka nepristranska (hitro bi lahko tudi direktno preverili, da res velja $\mathbb{E}(\hat{\theta}_{MM}) = \theta$). Na podoben način, kot smo to storili zgoraj, se dokopljemo do pomožnega rezultata, ki nam bo prav prišel kasneje. Velja:

$$\mathbb{E}(R^2) = \int_0^\infty r^2\ f(r \mid \theta)\ dr = 2\theta^2.$$

Sedaj se lotimo reševanja podnaloge c). Ker je cenilka za θ po metodi momentov nepristranska, velja $MSE(\hat{\theta}_{MM}) = Var(\hat{\theta}_{MM})$. Sedaj varianco tudi poračunajmo.

$$Var(\hat{\theta}_{MM}) = Var\left(\overline{R}\sqrt{\frac{2}{\pi}}\right) = \frac{2}{\pi}\frac{Var(R)}{n} = \frac{2}{\pi}\frac{\mathbb{E}(R^2) - (\mathbb{E}(R))^2}{n} =$$
$$= \frac{2}{\pi}\left(\frac{2\theta^2 - \left(\frac{\pi}{2}\theta\right)^2}{n}\right) = \frac{\theta^2}{n}\left(\frac{4-\pi}{\pi}\right)$$

Po zgornjem komentarju torej

$$MSE(\hat{\theta}_{MM}) = \frac{\theta^2}{n}\left(\frac{4-\pi}{\pi}\right) \approx \frac{\theta^2}{n}\cdot 0{,}273.$$

Pri izračunu asimptotične srednje kvadratične napake pri cenilki za θ po metodi največjega verjetja, nam na pomoč priskoči Fischerjeva informacija. Velja:

$$FI(\theta) = -\,\mathbb{E}\left(\frac{\partial^2 l(\theta \mid (r_1,\ \ldots,\ r_n))}{\partial\theta^2}\right) = -\,\mathbb{E}\left(\frac{1}{\theta^4}\left(2n\theta^2 - 3\sum_{i=1}^{n} r_i^2\right)\right) =$$
$$= -\,\mathbb{E}\left(\frac{1}{\theta^4}\left(2n\theta^2 - n\,\mathbb{E}(R^2)\right)\right) = -\,\mathbb{E}\left(\frac{1}{\theta^4}\left(2n\theta^2 - 3n\Big(\frac{\pi}{2}\theta^2 + \frac{4-\pi}{2}\theta^2\Big)\right)\right) = \frac{4n}{\theta^2},$$

oziroma:

$$FI^{-1}(\theta) = \frac{1}{4}\frac{\theta^2}{n},\ \text{zato asimptotično velja:}\ \ MSE(\hat{\theta}_{MNV}) \approx \frac{\theta^2}{n}\cdot 0{,}25.$$

Opazimo, da je faktor pred asimptotično srednje kvadratično napako za cenilko, dobljeno po metodi največjega verjetja nekoliko manjši, zato je vsaj asimptotično cenilka $\hat{\theta}_{MNV}$ nekoliko boljša od cenilke $\hat{\theta}_{MM}$.

3 Naloga 3

Naloga nam pove, da imamo podane mesečne temperature (torej 12 podatkov letno) za n zaporednih let (v našem primeru podatki med letoma 1986 in 2020, torej $n = 35$). Podnaloga a) nam narekuje, da podatke o meritvah zapišemo kot urejene pare, da se bomo lahko problema lotili z enostavno linearno regresijo. Opažene meritve so smiselno časovne urejene (leto-mesec), zato si podatke lahko predstavljamo kot pare $(x_i,\ y_i)_{i=1,\ 2,\ ...,\ 12n}$, kjer x_i predstavlja časovno komponento problema, y_i pa opaženo temperaturo pri časovno i-ti meritvi. V resnici lahko le pretvorimo podatek oblike: ($leto$, $mesec$, $temperatura$) v par ($leto$+$mesec$/12−1986, $temperatura$) (tu je −1986 le zato, ker je to leto našega prvega podatka, lahko bi brez tega). Sedaj označimo:

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_{12n} \end{bmatrix},\ Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{12n} \end{bmatrix} \text{ in } \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}.$$

V duhu enostavne linearne regresije iščemo β_0 in β_1, tako da se bo $X\beta$ po metodi najmanjših kvadratov najbolje prilegal Y. Teorija nam pove, da lahko β_0 in β_1 ocenimo s cenilkama:

$$\hat{\beta}_0 = \frac{\left(\sum_{i=1}^{12n} x_i^2\right)\left(\sum_{i=1}^{12n} y_i\right) - \left(\sum_{i=1}^{12n} x_i\right)\left(\sum_{i=1}^{12n} x_i y_i\right)}{12n\sum_{i=1}^{12n} x_i^2 - \left(\sum_{i=1}^{12n} x_i\right)^2}$$

$$\hat{\beta}_1 = \frac{12n\sum_{i=1}^{12n} x_i y_i - \left(\sum_{i=1}^{12n} x_i\right)\left(\sum_{i=1}^{12n} y_i\right)}{12n\sum_{i=1}^{12n} x_i^2 - \left(\sum_{i=1}^{12n} x_i\right)^2}$$

Literatura

[1] E. Zakrajšek, *Verižnica*, [ogled 30. 6. 2023], dostopno na `https://ucilnica.fmf.uni-lj.si/pluginfile.php/8283/mod_resource/conte4/predavanja/veriznica/veriznica.pdf`.

[2] M. Raič *Zapiski predavanj- Izražava povprečja in variance pri podatkih, organiziranih po skupinah* [ogled 3. 7. 2023], dostopno na `https://ucilnica.fmf.uni-lj.si/pluginfile.php/135261/mod_resource/content/2/Razcep_var.pdf`.