Projektna naloga iz statistike

Andraž Čepič

2. 6. 2022

V projektu ves čas uporabljamo Python s paketi Pandas, NumPy, Jupyter in Matplotlib. Vsi programi, spisani v namen obdelave podatkov, se nahajajo v mapi `skripte`.

Naloga 1

V namen obdelave podatkov smo napisali Jupyter zvezek `kibergrad.ipynb`. Za začetek naložimo podatke iz datoteke `kibergrad.csv` v Pandas DataFrame objekt.

(a)

Izberemo enostavni slučanji vzorec velikosti 200 s funkcijo `pandas.DataFrame.sample`. Če so

$$X_1, \ldots, X_{200}$$

števila otrok vsake od vzorčenih družin, je primerna ocena za povprečje enaka

$$\overline{X} = \frac{X_1 + \cdots + X_{200}}{200}.$$

Za naš specifičen vzorec dobimo oceno za povprečno število otrok v mestu Kibergrad:

$$\overline{X} = 0{,}925$$

(b)

Ocena za standardno napako je podana s formulo

$$\widehat{\mathrm{SE}}^2 = \frac{N-1}{N} \cdot \frac{1}{n(n-1)} \sum_{i=1}^{n} (X_i - \overline{X})^2,$$

kjer je N velikost populacije in n velikost enostavnega slučanega vzorca. V našem primeru je $N = 43.886$ in $n = 200$. Tako za naš vzorec dobimo

$$\widehat{\mathrm{SE}} = 0{,}0808$$

Za enostavno slučajno vzorčenje so intervali zaupanja ocen povprečja oblike

$$\overline{X} - \widehat{\mathrm{SE}} \cdot F_t^{-1}(1 - \frac{\alpha}{2}) < \mu < \overline{X} + \widehat{\mathrm{SE}} \cdot F_t^{-1}(1 - \frac{\alpha}{2}),$$

kjer je F_t komulativna funkcija Studentove t-porazdelitve z $n-1$ prostostnimi stopnjami in $\alpha = 0.05$ stopnja tveganja. V našem primeru dobimo interval zaupanja

$$0{,}7657 < \mu < 1{,}0843$$

(c)

Pravo populacijsko povprečje se glasi

$$\mu = \frac{x_1 + \ldots + x_N}{N} = 0{,}9479.$$

Prava standardna napaka za enostavni slučanji vzorec velikosti $n = 200$ je

$$\mathrm{SE}^2 = \frac{N-n}{N-1} \cdot \frac{\sigma^2}{n}, \tag{1}$$

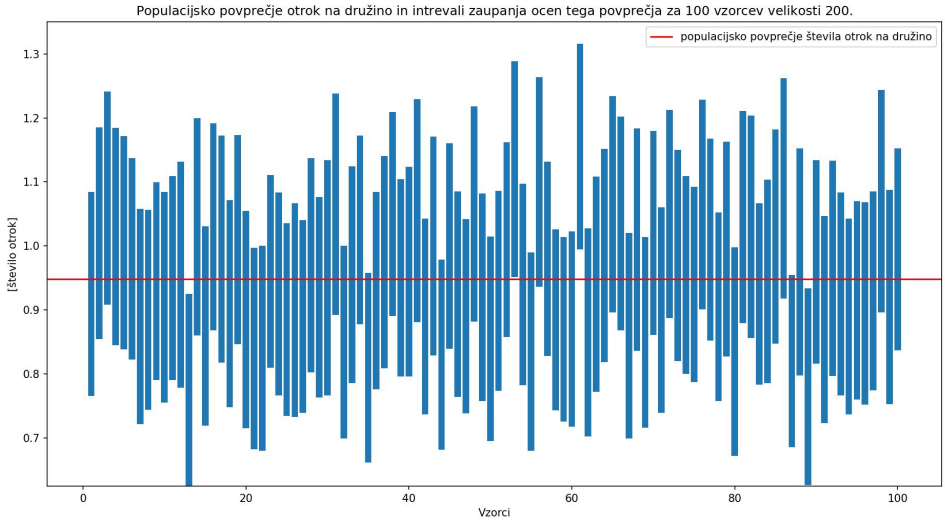
kjer je σ^2 variacija za celo populacijo. Za naše podatke je

$$\mathrm{SE} = 0{,}0816.$$

Opazimo, da je ocena za povprečje malce manjša od pravega povprečja in ocena za standardno napako je prav tako malo manjša, vendar se razlikuje šele v tretji decimalki. Da, interval zaupanja pokrije populacijsko povprečje.

(d)

Intervale zaupanja izračunamo na enak način, kot smo ga za prvi vzorec. Rezultati se nahajajo v mapi `rezultati`, in sicer v `intervali_zaupanja.csv`. Naslednja slika prikazuje te intervale zaupanja in populacijsko povprečje:



Izračunamo, da populacijsko povprečje pokrije 96 intervalov zaupanja, oz. delež intervalov, ki pokrijejo populacijsko povprečje, je 0,96.

(e)

Če označimo i-to oceno povprečja iz i-tega vzorca z μ_i in z $\overline{\mu}$ označimo povprečje teh ocen, je potem ocena za varianco teh ocen enaka

$$\widehat{\sigma}^2 = \frac{1}{100-1}\sum_{i=1}^{100}(\mu_i - \overline{\mu})^2.$$

Torej je standardni odklon enak

$$\widehat{\sigma} = 0{,}0776.$$

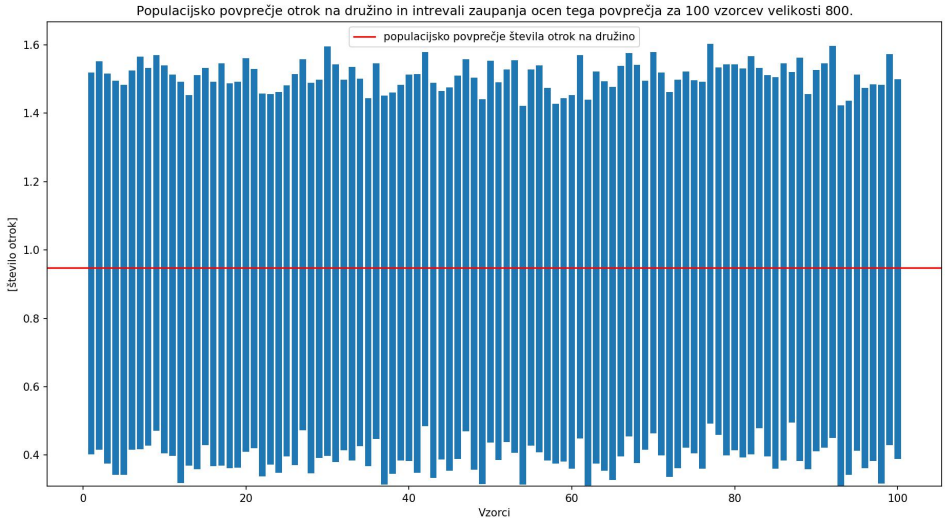
Prava standardna napaka za vzorec velikosti 200 pa je

$$\mathrm{SE} = 0{,}0816,$$

kar vemo že od prej. Opazimo, da je standardni odklon ocen povprečja manjši od standardne napake.

(f)

Tedaj so rezultati o intervalih zaupanja shranjeni v datoteki `intervali_zaupanja_1.csv` v mapi `rezultati`. Grafično je v tem primeru



Takoj opazimo, oziroma preverimo računsko, da tedaj vsi intervali zaupanja pokrijejo populacijsko povprečje.

V tem primeru je standardni odklon ocen za povprečje enak

$$\widehat{\sigma}_1 = 0{,}0399.$$

Prava standardna napaka za vzorec velikosti 800 pa je

$$\mathrm{SE}_1 = 0{,}0405.$$

Iz formule (1) je očitno, da je za večje velikosti vzorcev n standardna napaka manjša, zato nas ne preseneča, da je prava stantardna napaka za vzorce velikosti 800 precej manjša od tiste za vzorce velikosti 200. Zanimivo je, da se standardni odklon ocen za povprečje v obeh primerih precej ujema s pravimi napakami, torej je standardni odklon ocen za vzorce velikosti 800 približno pol manjši od tistega, ki pride iz vzorcev velikosti 200. TODO: pojasni do konca.

Naloga 2

Obdelava podatkov je v Jupyter datoteki `naloga_2.ipynb`.

(a)

Naj bo X spremenljivka z dano porazdelitivjo, odvisno od parametra θ. Tedaj je

$$\begin{aligned}\mathrm{E}(X) &= \frac{1}{3}\theta + \frac{4}{3}(1-\theta) + (1-\theta) \\ &= -2\theta + \frac{7}{3}.\end{aligned}$$

Torej je

$$\theta = \frac{7}{6} - \frac{1}{2}\,\mathrm{E}(X),$$

zato je

$$\widehat{\theta} := \frac{7}{6} - \frac{1}{2}\overline{X}$$

cenilka za parameter θ preko ocene prvega momenta $\overline{X}$. Ocena $\overline{X}$ za $\mathrm{E}(X)$ je nepristranska in pričakovana vrednost je linearna, torej je $\widehat{\theta}$ nepristranska cenilka za θ. Po naših opažanjih dobimo oceno

$$\widehat{\theta} = 0{,}3667.$$

(b)

Srednja kvadratična napaka te ocene je

$$\mathrm{SE}^2 = \mathrm{E}\left(\left(\widehat{\theta} - \theta\right)^2\right) = \mathrm{var}\left(\widehat{\theta}\right) = \frac{\sigma^2}{10},$$

zato je nepristranska ocena za kvadrat standardne napake pri metodi momentov enaka

$$\widehat{\mathrm{SE}}^2 = \frac{\widehat{\sigma}_+^2}{10} = \frac{1}{10} \cdot \frac{1}{10-1} \sum_{i=1}^{10} (X_i - \overline{X})^2$$

in pri teh konkretnih opažanjih ocena za standardno napako znaša

$$\widehat{\mathrm{SE}} = 0{,}3399.$$

(c)

Če so opažene vrednosti označene po vrsti z $x_1, \ldots, x_{10}$ in slučajne spremenljivke teh opažanj z $X_1, \ldots, X_{10}$, je verjetje za naša opažanja zaradi neodvisnosti enako

$$\begin{aligned} \mathrm{L}(\theta; X) &= \mathrm{P}(X_1 = x_1) \cdots \mathrm{P}(X_{10} = x_{10}) \\ &= \frac{2^6}{3^{10}} \cdot \theta^4 (1-\theta)^6. \end{aligned}$$

Iščemo maksimum verjetja. Lažje je opravljati z logaritmom, zato definiramo

$$l(\theta; X) = \log(\mathrm{L}(\theta; X)) = \log\left(\frac{2^6}{3^{10}}\right) + 4\log\theta + 6\log(1-\theta).$$

Logaritem je monotona funkcija, zato bo maksimum $\mathrm{L}(\theta; X)$ dosežen ob istem θ kot za $l(\theta; X)$. Odvod l je potem

$$\frac{\partial l}{\theta} = \frac{4}{\theta} - \frac{6}{1-\theta}.$$

Rešujemo torej enačbo

$$\frac{4}{\theta} - \frac{6}{1-\theta} = 0,$$

zato je

$$\begin{aligned} \frac{4}{\theta} &= \frac{6}{1-\theta} \\ 6\theta &= 4 - 4\theta \\ 5\theta &= 2 \\ \theta &= \frac{2}{5}. \end{aligned}$$

To je edini lokalni ekstrem na notranjosti intervala $[0, 1]$, hkrati pa je L na robovih intervala enak 0 in $L(\frac{2}{5}; X) > 0$, zato je to res globalni maksimum verjetja. Za oceno parametra θ po metodi največjega verjetja torej vzamemo

$$\widehat{\theta} = \frac{2}{5} = 0{,}4.$$

(d)

Za oceno standardne napake te ocene uporabimo Fisherjevo informacijo

$$\mathrm{FI}(\theta) = -\,\mathrm{E}\left(\frac{\partial^2 l}{\partial \theta^2}(\theta; X)\right).$$

Ocena za standardno napako se zdaj glasi

$$\widehat{\mathrm{SE}} = \frac{1}{\sqrt{\mathrm{FI}(\theta)}}.$$

V našem primeru je

$$\frac{\partial^2 l}{\partial \theta^2}(\theta; X) = -\frac{4}{\theta^2} - \frac{6}{(1-\theta)^2},$$

zato je

$$\mathrm{FI}(\theta) = \frac{4}{\theta^2} + \frac{6}{(1-\theta)^2}$$

in končno v našem primeru $\theta = \frac{2}{5}$ dobimo oceno za standardno napako:

$$\widehat{\mathrm{SE}} = 0{,}1549.$$

(e)

Naj bo f_θ gostota porazdelitve spremenljivke θ. Na intervalu $[0, 1]$ je potem konstantno enaka 1, drugje pa 0. Naj bo H dogodek, da so se zgodila neodvisna opažanja $X_1 = x_1, \ldots, X_{10} = x_{10}$ kot podano. Za pogojno gostoto $f_{\theta|H}$ ob opaženem dogodku H uporabimo Bayesovo formulo za mešane porazdelitve:

$$f_{\theta|H}(t) = \frac{\mathrm{P}(H|\theta = t) \cdot f_\theta(t)}{\mathrm{P}(H)},$$

kjer je verjetnost dogodka H enaka

$$\mathrm{P}(H) = \int_0^1 \mathrm{P}(H|\theta = t) f_\theta(t) dt.$$

Vemo, da je

$$\mathrm{P}(H|\theta = t) = \frac{2^6}{3^{10}} t^4 (1-t)^6,$$

torej je

$$\mathrm{P}(H) = \int_0^1 \frac{2^6}{3^{10}} t^4 (1-t)^6 dt = \frac{2^6}{3^{10}} B(5, 7).$$

Sklepamo, da je

$$f_{\theta|H}(t) = \frac{1}{B(5,7)} t^4 (1-t)^6,$$

kjer je $t \in [0, 1]$, drugje je pa enaka 0. Ugotovili smo, da je aposteriorna porazdelitev $\theta|H$ porazdeljena z beta porazdelitvijo $B(5, 7)$.