Projektna naloga iz statistike

Andraž Čepič

2. 6. 2022

V projektu ves čas uporabljamo Python s paketi Pandas, NumPy, Jupyter in Matplotlib.

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 200 je

$$\mathrm{SE}^2 = \frac{N - 200}{N - 1} \cdot \frac{\sigma^2}{200},$$

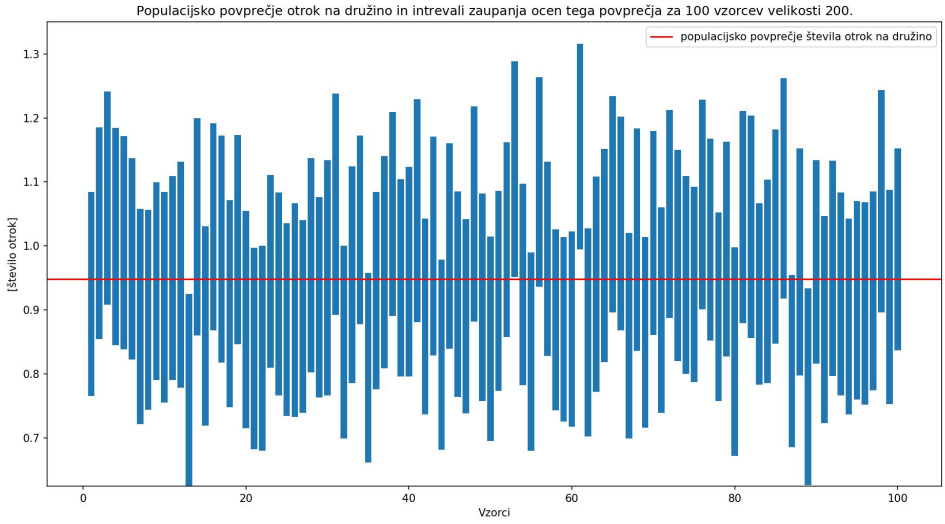
kjer je σ^2 variacija za celo populacijo. Za naše podatke je

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

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

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

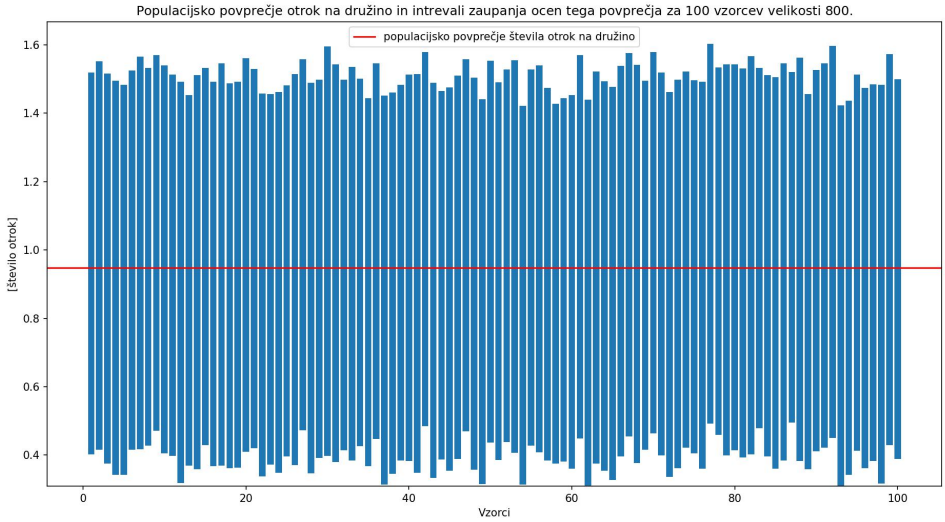
Prava standardna napaka za vzorec velikosti 200 pa je

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

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 intervali zaupanja vsi 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

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