Seminarska naloga iz Statistike

Dejan Perić

25. december 2021

1 Prva naloga

V datoteki Kibergrad se nahajajo informacije o 43.886 družinah, ki stanujejo v mestu Kibergrad. Za vsako družino so zabeleženi naslednji podatki (ne boste potrebovali vseh):

- Tip družine (od 1 do 3)
- Število članov družine
- Število otrok v družini
- Skupni dohodek družine
- Mestna četrt, v kateri stanuje družina (od 1 do 4)
- Stopnja izobrazbe vodje gospodinjstva (od 31 do 46)

1.1 a)

Izmed 43886 družin izberemo vzorec družin velikosti 200. Vzorce bomo pridobivali s pomočjo sample iz knjižnice random. Za cenilko povprečnega števila otrok bomo vzeli cenilko

$$\hat{\mu} = \frac{1}{200} \sum_{i=1}^{200} X_i$$

Dobimo oceno, da je povprečno število otrok v Kibergradu enako $0,975$.

1.2 b)

Za cenilko standardne napake bomo vzeli cenilko iz predavanj. Ta je enaka

$$\hat{SE}_+^2 = \frac{N-n}{N-1} \cdot \frac{\hat{\sigma}_+^2}{n} \ ,$$

pri čemer je $\hat{\sigma}_+^2$ enaka

$$\hat{\sigma}_+^2 = \frac{N-1}{N} \cdot \frac{1}{n-1} \sum_{i=1}^{n} (X_i - X)^2 \ .$$

$$\hat{SE}_+^2 = \frac{N-n}{N} \cdot \frac{1}{n \cdot (n-1)} \sum_{i=1}^{n} (X_i - X)^2$$

Pri nas je $N = 43886$ in $n = 200$, torej je

$$\hat{SE}_+^2 = \frac{43686}{43886} \cdot \frac{1}{200 \cdot 199} \sum_{i=1}^{200} (X_i - X)^2 \ .$$

Izračunamo, da je standardna napaka enaka 0.09054913965779846.
Sedaj s pomočjo studentove porazdelitve izračunajmo eksakten 95% interval zaupanja. Iz predavanj vemo, da je ta oblike

$$\mu \in (\hat{\mu} - \hat{SE}_+ \cdot F^{-1}_{Student(n-1)}(1 - \frac{\alpha}{2}), \hat{\mu} + \hat{SE}_+ \cdot F^{-1}_{Student(n-1)}(1 - \frac{\alpha}{2})) \ .$$

Oceni za povprečno število otrok ($\hat{\mu}$) in standardno napako ($\hat{SE}_+$) že imamo. Stopnja tveganja je enaka $\alpha = 0,05$. S pomočjo knjižnice *scipy.stats* izračunamo, da je

$$F^{-1}_{Student(n-1)}(0,975) = 1.9719565442493954$$

Interval zaupanja je torej enak $(0.7379511969254118, 1.0620488030745883)$

1.3 c)

Vzorčno povprečje in ocenjeno standardno napako primerjajte s populacijskim povprečjem in pravo standardno napako. Ali interval zaupanja iz prejšnje točke pokrije populacijsko povprečje?
Za cenilko povprečja celotne populacije bomo vzeli isto kot smo za oceno populacijskega povprečja pri vzorcu.

$$\mu = \frac{1}{43886} \sum_{i=1}^{200} X_i = 0.9479332816843641$$

Opazimo, da se populacijsko povprečje in vzorčno povprečje ujemata zgolj v eni decimalki. Iz vseh podatkov družin lahko izračunamo varianco števila otrok družin, s pomočjo katere bomo lahko ocenili kakšna je prava standardna napaka za vzorec velikosti 200. Ker imamo podatke celotne populacije, za cenilko variance vzamemo

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{43886} (X_i - X)^2 = 1.3391064929282408$$

Za cenilko prave standardne napake vzamemo cenilko, najdeno v knjigi ????:

$$SE^2 = \frac{\sigma^2}{n} \cdot (\frac{N-n}{N-1}) = \frac{\sigma^2}{200} \cdot \frac{43686}{43885}$$

Dobimo, da je prava standardna napaka ocenjena z $SE = 0.0816404987959038$. Absolutna napaka ocenjene standardne napake je torej ???, relativna pa ???. Opazimo, da je

$$\mu = 0.9479332816843641 \in (0.7379511969254118, 1.0620488030745883) \ ,$$

torej interval zaupanja iz prejšnje točke pokrije populacijsko povprečje.

1.4 d)

Vzemite še 99 enostavnih slučajnih vzorcev in prav tako za vsakega določite 95% interval zaupanja. Narišite intervale zaupanja, ki pripadajo tem 100 vzorcem. Koliko jih pokrije populacijsko povprečje?

1.5 e)

Izračunajte standardni odklon vzorčnih povprečij za 100 prej dobljenih vzorcev. Primerjajte s pravo standardno napako za vzorec velikosti 200.

1.6 f)

Izvedite prejšnji dve točki še na 100 vzorcih po 800 družin. Primerjajte in razložite razlike s teorijo vzorčenja.

2 Druga naloga

2.1 a)

Ocenite povprečje in standardni odklon za telesno temperaturo posebej pri moških in posebej pri ženskah.

2.2 b)

Za povprečji iz prejšnje točke določite 95% intervala zaupanja.

2.3 c)

Preizkusite domnevo, da imajo moški in ženske v povprečju enako telesno temperaturo.

3 Tretja naloga

V datoteki Temp LJ se nahajajo izmerjene mesečne temperature v Ljubljani v letih od 1986 do 2020. Postavimo naslednja dva modela spreminjanja temperature s časom:

- Model A: vključuje linearni trend in sinusno nihanje s periodo eno leto.
- Model B: vključuje linearni trend in spreminjanje temperature za vsak mesec posebej.

Očitno je model B širši od modela A.

3.1 a)

Preizkusite model A znotraj modela B.

3.2 b)

Pri modeliranju je nevarno privzeti preširok model: lahko bi recimo postavili model, po katerem je temperatura vsak mesec drugačna, neidvisno od ostalih mesecev, a tak model bi bil neuporaben za napovedovanje. Akaikejeva informacija nam pomaga poiskati optimalni model – izberemo tistega, za katerega je le-ta najmanjša. Akaikejeva informacija je sicer definirana z verjetjem, a pri linearni regresiji in Gaussovem modelu je le-ta ekvivalentna naslednji modifikaciji:

$$\mathrm{AIC} := 2m + n \ln \mathrm{RSS},$$

kjer je m število parametrov, n pa je število opažanj. Kateri od zgornjih dveh modelov ima manjšo Akaikejevo informacijo?

4 Literatura