Projetkna naloga pri statistiki

Tadej Mohorčič

Python datoteke z rešenimi nalogami lahko najedete na GitHubu: GitHub repozitorij

1 Prva naloga

Pri tej nalogi smo obravavali 43886 družin, ki živijo v mestu Kibergrad. Natančneje, morali smo obravnavati delež družin, kjer glava družine nima srednješoljske izobrazbe, to pa bomo storili z enostavnim vzorčenjem in ocenjevanjem. Med podatke smo dodali spremenljivko, poimenovano **Brez Izobrazbe**. Ta je indikator dogodka, ali imamo srednješolsko izobrazbo ali ne, torej ima vrednosti le 0 in 1. Tukaj smo vrednost 1 dodelili tistim družinam, kjer glava nima srednješoljske izobrazbe. Nekatere podatke sem zbral v tabeli Data200 in Data800.

1.1 A del naloge

Tukaj smo vzeli naključno izbran delež 200 družin iz celotne populacije. Kot oceno za delež populacije smo tako vzeli kar delež na tem naključno izbranem vzorcu.

Ocena za delež je 0.205**.**

1.2 B del naloge

Oceno za standardno napako najdemo s formulo

$$\widehat{\mathrm{SE}}_{+}^{2} = \sqrt{\frac{(N-n)\,\mathrm{p}(1-\mathrm{p})}{(n-1)N}}$$

kjer je N velikost populacije, n velikost vzorca, p pa je populacijski delež. Za p tukaj uporabimo rezultat iz prejšnje naloge. Za interval zaupanja pa lahko vzamemo kar $\mathrm{p} \pm 1.96\widehat{\mathrm{SE}}_{+}^{2}$.

Standardna napaka je 0.0299879137**.**

Interval zaupanja je $(0.1671131677, 0.2992038693)$**.**

1.3 C del naloge

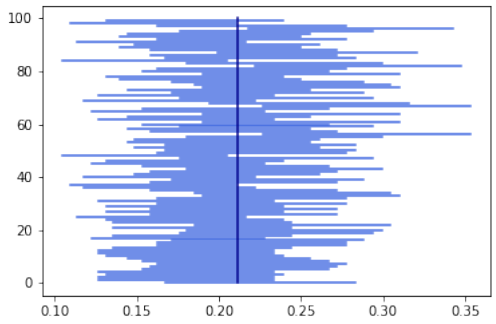
Pravi populacijski delež in pravo standardno napako najdemo kar z uporabo funkcij .mean() in .var().

Pravi populacijski delež je 0.2115025293**.**

Prava standardna napaka je 0.0288767216**.**

Interval zaupanja pokrije pravi populacijski delež.

1.4 D del naloge



Slika 1: IZ za 100 vzorcev velikosti 200 družin

Tukaj smo morali zgenerirati še 99 vzorcev velikosti 200, in za vsakega izračunati interval zaupanja, ti pa so prikazani na sliki 1.

92 **intervalov zaupanja pokrije populacijski delež.**

1.5 E del naloge

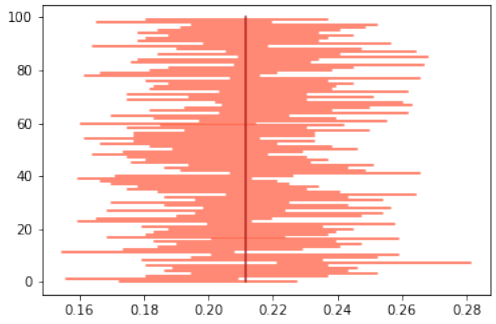
Standardni odklon vzorčnih deležev izračunamo kar z .std(), standardno napako za vzorec velikosti 200 pa po enaki formuli kot prej, le da vstavimo v formulo populacijski delež. Pri primerjavi pride do zelo majhnega odstopanja, zato bi rekli da je standardni odklon primerljiv z standardno napako za vzorec velikosti 200.

Standardni odklon je 0.0297648702**.**

Standardna napaka za vzorec velikosti 200 **je** 0.0288828163**.**

1.6 F del naloge

Zadnji dve točki smo morali ponoviti še za vzorce velikosti 800. Do rezultatov pridemo po isti poti kot zgoraj.



Slika 2: IZ za 100 vzorcev velikosti 800 družin

Intervali zaupanja so za večji populacijski delež precej bolj skoncentrirani ob sredini, prav tako pa so tudi ožji, kar lahko razberemo iz tabele Data800. Opazno pa je tudi standardni odklon precej manjši, in tudi bližji pravi standardni napaki za vzorev velikosti 800. Verjetno se bosta ti dve vrednosti še bolj zbližali, če velikost vzorca pošljemo proti večji številki.

97 **intervalov zaupanja pokrije populacijski delež.**

Standardni odklon je 0.0141844710

Standardna napaka za vzorec velikosti 800 **je** 0.0143149434**.**

2 Druga naloga

Pri tej nalogi smo obdelali odčitane telesne temperature pri moških in pri ženskah, kjer smo privzeli, da sta za oba spola ti vrednosti porazdeljeni normalno.

2.1 A del naloge

Za oceno povprečja in standardnega odklon lahko uporabimo kar funkciji .mean() in .std(). Vemo, da je povprečje na vzorcu dobra cenilka za povprečje, medtem ko funckija .std() zraven še upošteva faktor $\frac{1}{N-1}$, za katerega pa tudi vemo, da je nepristranska cenilka.

Povprečje je 36.7247863248 **pri moških in** 36.8854700855 **pri ženskah.**

Standardni odklon je 0.3881976457 **pri moških in** 0.4130487515 **pri ženskah.**

2.2 B del naloge

Za obe povprečji smo morali določiti 95% interval zaupanja. V šoli smo pokazali, da če σ ni znan (pri nas ne poznamo točne vrednosti), potem za X, ki je porazdeljen normalno, velja

$$\frac{\overline{X} - \mu}{\widehat{\sigma}^+}\sqrt{n} \sim Student(n-1)$$

Tako bomo tukaj morali namesto funkcije napake erf uporabiti funkcijo Studentove porazdelitve. Kalkulator na spletu nam vrne, da je za 95% zaupanja pri 64 prostorskih stopnjah vrednosti 1.9977. Dobimo, da je

$$\mathrm{C}_\alpha = 1.9977\frac{\widehat{\sigma}^+}{\sqrt{n}}$$

Interval zaupanja za moške je $(36.6285970859, 36.8209755637)$**.**

Interval zaupanja za ženske je $(36.7831231354, 36.9878170355)$**.**

2.3 C del naloge

Sedaj pa moramo testirati domnevo, da imajo moški in ženske v povprečju enako telesno temperaturo. Imamo torej $M_1, \ldots, M_{65}$ in $Z_1, \ldots, Z_{65}$ porazdeljeno normalno z μ_M in μ_Z ter testiramo:

- H_0: $\mu_M = \mu_Z$
- H_1: $\mu_M \neq \mu_Z$

Na vajah smo pokazali, da za spremenljivko T velja

$$T \coloneqq \frac{\overline{X} - \overline{Y}}{S}\sqrt{\frac{mn}{m+n}} \sim Student(n+m-2),$$

kjer je

$$S \coloneqq \sqrt{\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2 + \sum_{i=1}^{m}(Y_i - \overline{Y})^2}{m+n-2}}.$$

Pri nas je $m = n = 65$, X in Y pa sta ravno odčitane vrednosti temperature moških in žensk. Z uporabo formule

$$P(|\overline{M} - \overline{Z}| \leq \mathrm{C}_\alpha) \leq \alpha$$

dobimo rezultate. Ker je $|\overline{M} - \overline{Z}| = 0.1606837607$, bomo domnevo obdržali le pri 0.01 stopnji tveganja.

Pri stopnji tveganja 0.05 **je interval** $(-0.1391179454, 0.1391179454)$**.**

Pri stopnji tveganja 0.01 **je interval** $(-0.1838407053, 0.1838407053)$**.**

3 Tretja naloga

Pri tej nalogi pa se ukvarjamo z izmerjenimi temperaturami v Ljubljani med leti 1986 ter 2020.

3.1 A del naloge

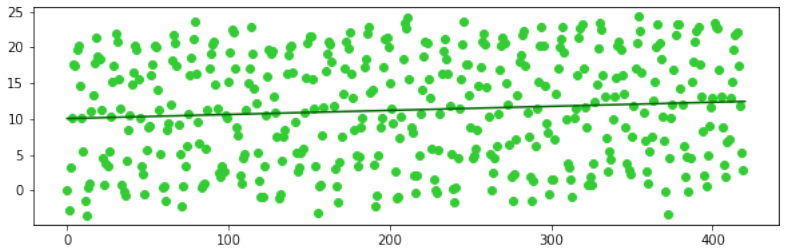
Najprej smo morali modelirati spreminjanje temperature z enostavno linearno regresijo, torej poiskati takšne koeficiente a in b, da se bo premica

$$y = ax + b$$

najbolj prilegala našim meritvam. Vektor enic ter vektor x-ov lahko zapakiramo v 420×2 matriko, ki izgleda tako:

$$\begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ \vdots & \vdots \\ 1 & 418 \\ 1 & 419 \end{bmatrix}$$

Predoločen sistem $y = A * x$ lahko rešimo z metodo najmanjših kvadratov: $x = (A^\intercal A)^{-1} A^\intercal y$, lahko pa uporabimo tudi katero od funkcij iz paketa scipy.stats.



Slika 3: Enostavna linearna regresija

Opazimo lahko, da se letno temperatura poveča za 0.06830 stopinje Celzij. Prav tako lahko poračunamo p-vrednost, ki je enaka 0.06221, kar pomeni, da bomo obdržali ničelno hipotezo.

Linerni trend spreminjanja temperature ni statistično značilen.

3.2 B del naloge

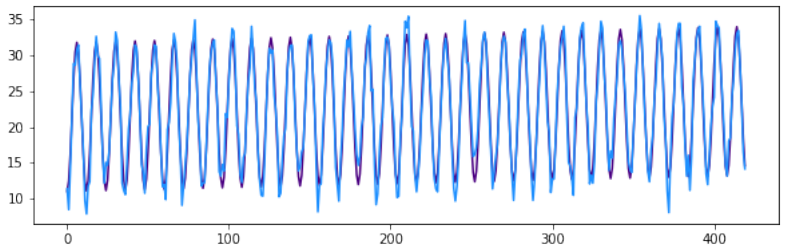
Sedaj pa smo morali modelirati spreminjanje temperature tako, da smo zraven upoštevali še periodično nihanje temperature. Uporabil bom periodo 1 leto, saj se mi zdi, da se letno zgodita ekstrema le v poletnih in zimskih mesecih. Radi bi našli torej koeficiente a, b, c in d, da se bo

$$y = a + bx + c\sin(x) + d\cos(x)$$

najbolj prilegala našim točkam. Tako smo dobili matriko

$$\begin{bmatrix} 1 & 0 & 0 & 1 \\ 1 & 1 & \sin(\frac{\pi}{6}) & \cos(\frac{\pi}{6}) \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 419 & \sin(\frac{419\pi}{6}) & \cos(\frac{419\pi}{6}) \end{bmatrix}$$

Sedaj imamo spet predoločen sistem, tokrat velikosti 420×4.



Slika 4: Linearna regresija s periodo nihanja 1 leto

Tudi tokrat pride do letnega spreminjanja temperature, letno pa se temperatura poveča za 0.06422.

3.3 C del naloge

Tukaj smo morali napovedati povprečno letno temperaturo za 2040 ter za mesec Januar istega leta. Zato lahko uporabimo kar krivuljo ali premico iz A ali B dela naloge. Za 95% interval zaupanja pa uporabimo formulo

$$\hat{y} \pm t_{crit} * s_{yx} \sqrt{\frac{1}{n} + \frac{(x - \overline{x})}{SS_x}}$$

kjer je $\hat{y}$ napovedana vrednost, x vrednost, pri kateri napovedujemo, t_{crit} pa t-vrednost, ki jo dobimo iz stopnje tveganja ter stopnje prostosti.

Napovedana temperatura za mesec Januar 2040 je $3.21778426°C$**, interval zaupanja pa je** $(0.49553924°C, 5.940029°C)$**.**

Napovedana povprečna temperatura za leto 2040 je $13.63297137°C$**, interval zaupanja pa je** $(9.96553995°C, 17.60218744°C)$**.**